FACULTEIT
BIO-INGENIEURSWETENSCHAPPEN

GEBRUIK VAN MACHINE LEARNING METHODEN VOOR HET VOORSPELLEN VAN PRODUCTIEKENGETALLEN BIJ MELKVEE

Aantal woorden: 34.062

Maarten Perneel

Studentennummer: 01503165

Promotor(en):

Prof. dr. ir. Stefaan De Smet

Prof. dr. ir. Jan Verwaeren

Masterproef voorgelegd voor het behalen van de graad master in de Bio-ingenieurswetenschappen: landbouwkunde.

Academiejaar: 2019 - 2020

UNIVERSITEIT
GENT

DANKWOORD

Deze masterproef rond het gebruik van machine learning methoden voor het voorspellen van productiekengetallen bij melkvee, vormt het sluitstuk van mijn studies in de bio-ingenieurswetenschappen: landbouwkunde. Bij het verwezenlijken van deze masterproef kon ik rekenen op de steun en medewerking van verschillende personen, die ik hier wens te bedanken.

Als eerste wil ik mijn beide promotoren bedanken. Voor al mijn vragen rond machine learning kon ik steeds rekenen op prof. dr. ir. Jan Verwaeren, terwijl prof. dr. ir. Stefaan De Smet mij steeds kon helpen bij meer landbouwgerichte vragen. Ze vulden elkaar dan ook perfect aan en waren essentieel om deze masterproef tot een goed einde te brengen.

Daarnaast wens ik ook Coöperatie CRV uitvoerig te bedanken, in het bijzonder mijn interne begeleiders René van der Linde en Mikael Bastian. Zonder de data die CRV ter beschikking heeft gesteld, was het immers onmogelijk geweest om deze masterproef te realiseren.

Tot slot wil ik ook mijn ouders, broer en zus bedanken voor de steun die zij mij hebben verleend tijdens mijn studies.

Inhoudsopgave

SAMENVATTING

In deze masterproef werd onderzocht welke meerwaarde machine learning methoden kunnen bieden voor het voorspellen van productiekengetallen bij melkvee. Om de nodige achtergrondkennis te verwerven, werd eerst een literatuuronderzoek uitgevoerd waarbij informatie werd verzameld over welke factoren allemaal invloed hebben op de productieresultaten van melkkoeien (genetica, epigenetica, milieu) en hoe machine learning methoden kunnen worden gebruikt bij het berekenen van fokwaarden.

Vervolgens werden de door CRV aangeleverde ruwe data onder een passende vorm gebracht, teneinde de data maximaal te kunnen benutten bij het opstellen van machine learning modellen. Hierbij werd onder andere (1) gebruik gemaakt van lactatiecurves om de grote variabiliteit in hoeveelheid beschikbare mpr-metingen te kunnen weergeven in een beperkt aantal variabelen met een eenduidige betekenis en werd (2) aandacht besteed aan hoe een interactie-effect tussen de fokwaarden en de bedrijfsomstandigheden in rekening gebracht kon worden, zonder dat hierbij de beperking ontstond dat de opgestelde modellen enkel maar konden toegepast worden op bedrijven die waren opgenomen in de trainingsdataset.

Nadat de ruwe data onder een bruikbare vorm waren gebracht, werden meer dan 10 machine learning technieken toegepast op de data, gaande van klassieke multiple lineare regressie, over random forest regressie, tot support vector regressie met een radiale kernel. Hierbij werden 16 verschillende (productie)kengetallen beschouwd en werden er modellen opgesteld op verschillende momenten tijdens het leven van de dieren, beginnende vanaf de geboorte tot op een leeftijd van negen jaar. Daarnaast werden de machine learning technieken ook via drie verschillende strategieën toegepast: (1) één model op landelijk niveau, (2) voor ieder bedrijf een bedrijfsspecifiek model en (3) stacking van verschillende machine learning technieken. Dit resulteerde in meer dan 15000 verschillende combinaties van een machine learning techniek, een kengetal, een tijdstip en een strategie. Al deze combinaties werden een na een uitgewerkt en geëvalueerd, waarna de resultaten werden samengevat in overzichtelijke figuren en tabellen.

Uit deze resultaten is gebleken dat, indien er een voldoende hoog aantal waarnemingen is, de meerwaarde van individuele machine learning technieken relatief beperkt

is. Machine learning technieken die in staat zijn om interacties tussen de verschillende variabelen in de dataset in rekening te brengen, zoals support vector regressie met een radiale kernel, realiseren doorgaans een iets hogere performantie dan multiple lineaire regressie, maar de meerwaarde blijft steeds beperkt. Indien de verschillende beschouwde individuele machine learning technieken echter worden gecombineerd via stacking, bieden machine learning technieken wél een aanzienlijke meerwaarde. Bijvoorbeeld bij het productiekengetal levensproductie, uitgedrukt in kg FPCM, kon via stacking een R^2-waarde gerealiseerd worden van 0.47, een stijging van maar liefst 0.17 tegenover de R^2-waarden van het best presterende individuele machine learning model.

Bij een beperkte hoeveelheid waarnemingen, indien men bijvoorbeeld voor een individueel bedrijf een predictief model wil opstellen, bleken lasso regressie en random forest regressie de meest geschikte machine learning methoden te zijn om productiekengetallen bij melkvee te voorspellen. Doordat deze technieken over de mogelijkheid beschikken om predictieve modellen op te stellen op basis van datasets met meer variabelen dan waarnemingen, bieden deze machine learning technieken bovendien een groot voordeel tegenover multiple lineaire regressie. De modelperformantie die kan gerealiseerd worden met een bedrijfsspecifiek machine learning model is als gevolg van de beperkte hoeveelheid waarnemingen per bedrijf echter steeds lager dan deze die behaald kan worden met landelijke modellen, zelfs al gebruikt men machine learning technieken zoals random forest regressie of lasso regressie.

Ondanks het feit dat een passende primaire dataverwerking en stacking van machine learning technieken voor de meeste productiekengetallen resulteerde in R^2-waarden van ±0.50 of hoger, bleven de predictiefouten echter meestal erg groot. De relatief grote predictiefouten staan echter geen praktijktoepassingen van de opgestelde machine learning modellen in de weg, zo kon aangetoond worden dat machine learning modellen een grote meerwaarde kunnen bieden bij het afvoeren van overtollig jongvee met het oog op een verhoogde levensproductie. In vergelijking met de selectierespons die hierbij kan worden behaald met multiple lineaire regressie op de fokwaarden, slaagde het beste machine learning model er zelfs in om deze selectierespons te verdubbelen.

SUMMARY

In this master thesis, the possible advantages of using machine learning methods to predict production key performance indicators of dairy cattle, were explored. First a scientific literature review was carried out to gain insight in factors which influence milk production levels of dairy cows (genetics, epigenetics, and environment) and how machine learning methods could be used during breeding value estimation.

Thereafter, the raw data, delivered by CRV, where pre-processed with the aim of maximising information-use efficiency during the training of machine learning models. During the pre-processing, there was, amongst other things, (1) made use of lactation curves to be able to represent the great variability in the number of available milk production measurements in a restricted number of well-defined variables and (2) there was paid attention to how there could be accounted for an interaction effect between the breeding values and the farm circumstances, without creating the restriction that the trained machine learning models could only be applied on farms which were present in the training dataset.

After the raw data were pre-processed in an appropriate way, more than 10 machine learning methods were applied onto the data, from classic multiple linear regression, over random forest regression, until support vector regression with a radial kernel. Hereby, 16 different (production) key performance indicators were considered and machine learning models were trained for different moments during the life of dairy cattle, starting from birth until the age of nine years. In addition, the machine learning methods were applied following three different strategies: (1) one model on a national level, (2) for each farm a farm-specific model and (3) stacking of different machine learning methods. This resulted in more than 15000 different combinations of a machine learning method, a key performance indicator, a time moment and a strategy. All these combinations where, one after another, elaborated and evaluated, after which the results were represented in a number of clear figures and tables.

The obtained results have shown that, if there is a sufficiently high number of records available, the added value of individual machine learning methods is relatively restricted. Machine learning methods which are capable to account for interaction-effects between the different variables in the dataset, like support vector regression with a

radial kernel, often realise a slightly higher performance compared to multiple linear regression, but the difference is always relatively restricted. However, if all the considered machine learning methods were combined with stacking, the use of machine learning methods resulted in a considerable increase of the predictive performance. For example, for the lifetime production, expressed in kg FPCM, stacking of different machine learning models resulted in an R^2-score of 0.47, which is an increment of no less than 0.17 compared to the R^2-score of the best performing individual machine learning model.

If the number of records is relatively restricted, for example if one wants to set up a predictive model for an individual farm, lasso regression and random forest regression seemed the most appropriate machine learning methods to predict production key performance indicators of dairy cattle. Due to the fact these techniques have the possibility to generate a model based upon datasets with less observations compared to variables, these machine learning methods offer a great advantage compared to multiple linear regression. However, due to the restricted number of records which is often available on a single farm, the predictive performance which can be realised is always lower compared to the performance which can be obtained with models set up on a national level, even if one uses machine learning techniques like random forest regression or lasso regression.

Despite the fact that an appropriate data pre-processing and stacking of machine learning models resulted for most key performance indicators in values for R^2 of ±0.50 or higher, the prediction errors often stayed quite high. However, the relatively large prediction errors didn't seemed to prevent practical applications of the created machine learning models. For example, there could be shown that machine learning models can result in a considerable better selection of surplus youngstock to sell. If the aim of selection is an increased lifetime production, the best machine learning model could even double the selection response which can be realised with multiple linear regression on the breeding values.

Lijst van figuren

Lijst van tabellen

AFKORTINGEN

μ	gemiddelde.
σ^2	variantie.
%E	percentage Eiwit.
%V	percentage Vet.
A	Additief genetisch effect Additief genetische verwantschapsmatrix, pedigree verwantschapsmatrix.
BLUP	Best Linear Unbiased Prediction.
BV	Breeding Value, fokwaarde.
D	Dominant genetisch effect.
dgn	dagen.
E	Environment, omgeving.
EVA	Eén model Voor Allen, één model op landelijk niveau.
EVI	Eén model Voor Ieder, één model voor ieder bedrijf.
FCM	Fat Corrected Milk.
FPCM	Fat and Protein Corrected Milk.
G	Genotype.
GBV	Genomic Breeding Value.
I	epistatisch genetisch effect.
ins	inseminatie.
kg E	kg Eiwit.
kg M	kg Melk.

kg V kg Vet.

kg V+E kg Vet + Eiwit.

lasso lasso regressie.

LEK Leeftijd bij Eerste Kalving.

lft leeftijd.

linreg lineaire regressie.

lw lactatiewaarde.

M epigenetisch effect.

m.b.t. met betrekking tot.

m.u.v. met uitzondering van.

MBV Molecular Breeding Value.

mpr melkproductieregistratie.

MSE Mean Squared Error, gemiddelde kwadratische fout.

ncRNA non coding RNA, niet coderend RNA.

NN Neuraal Netwerk.

P Phenotype, fenotype.

PBV Pedigree Breeding Value.

PCA Principale Componenten Analyse.

QTL Quantitative Trait Loci.

r correlatie.

RH Relative Humidity, relatieve luchtvochtigheid.

ridge ridge regressie.

SNP Single Nucleotide Polymorphism.

ssGBLUP single step Genomic Best Linear Unbiased Prediction.

std standaardafwijking.

STIN STackINg.

SVR Support Vector Regressie.

THI Temperature Humidity Index, temperatuur-luchtvochtigheid index,.

tkt tussenkalftijd.

UBN Unieke BedrijfsNummer.

vgl vergelijking.

WISKUNDIGE CONVENTIES

a	constante
$\boldsymbol{a}$	vector
A	matrix
$\lvert a \rvert$	absolute waarde van a
Δa	verandering van a
$\hat{\boldsymbol{a}}$	schatter van $\boldsymbol{a}$
$\overline{\boldsymbol{a}}$	gemiddelde van $\boldsymbol{a}$
$\lVert \boldsymbol{a} \rVert$	norm van $\boldsymbol{a}$
I	eenheidsmatrix
I_n	$n \times n$ eenheidsmatrix
$A^{(n \times n)}$	$n \times n$ matrix
A^{-1}	inverse van A
$A^{\top}$	getransponeerde van A
$E[X]$	verwachtingswaarde van X
$Var(X)$	variantie van X
$\sim N(\mu, \sigma^2)$	normaal verdeeld met gemiddelde μ en variantie σ^2

HOOFDSTUK 1

LITERATUURSTUDIE

De meeste productie-eigenschappen van melkvee worden niet enkel door het genotype bepaald, maar staan ook onder invloed van de omgeving waarin het dier leeft. Om de invloed van het genotype (G) en de omgeving (E) op het fenotype (P) van een dier te verklaren, wordt daarom vaak model (1.1) gebruikt.

$$\begin{aligned} P &= G + E + G \times E \\ G &= A + D + I \end{aligned} \tag{1.1}$$

met

- A: additief genetisch effect, effect van individuele allelen
- D: dominant genetisch effect, interactie-effect tussen allelen op eenzelfde locus
- I: epistatisch genetisch effect, interactie-effect tussen allelen op verschillende loci
- E: omgevingseffect

Van de vier factoren die in model (1.1) worden onderscheiden, zijn er 3 genetische factoren (A, D, I) en 1 niet-genetische factor (E). Daarnaast is er nog epigenetica (M), die een vijfde factor vormt die invloed heeft op het fenotype, maar niet in rekening wordt gebracht bij model (1.1), ondanks dat uit de literatuur blijkt dat epigenetica een significante invloed heeft op bijvoorbeeld de melkproductie in eerste lactatie (González-Recio et al., 2012). De reden voor de afwezighed van M in model (1.1), is dat epigenetica zowel tot G als tot E behoort (zie §1.4) waardoor epigenetica moeilijk in model (1.1) is in te passen.

Betreffende de invloedsfactoren op het fenotype die gelinkt zijn aan de DNA-sequentie (A, D, I) wordt er in het kader van de veredeling voornamelijk gefocust op het berekenen van het additief genetisch effect A, ook wel de fokwaarde (*breeding value , BV*) genoemd. A is immers de enige factor die gelinkt is aan de DNA-sequentie en op relatief voorspelbare wijze[1] overgedragen wordt van stier/koe naar kalf. Dit in

[1] $E[A_{kalf}] = 0.5\,A_{koe} + 0.5\,A_{stier}$

tegenstelling tot de andere componenten van G, die niet (D) of slechts beperkt en onvoorspelbaar (I) overdraagbaar zijn van stier/koe naar kalf door de processen die eigen zijn aan de meiose: (1) recombinatie van het genetisch materiaal door crossovers en (2) halvering van het chromosomenaantal.

In het kader van deze masterproef, waarbij getracht wordt om de productiekengetallen die individuele dieren zullen realiseren (P) te voorspellen, zijn echter alle invloedsfactoren op het fenotype (model (1.1)+ M), relevant. Daarom wordt in deze literatuurstudie dieper ingegaan op welke modellen er bestaan om fokwaardes te schatten (§1.1, §1.2) en wat er geweten is omtrent omgevingsfactoren (§1.3) en epigenetica (§1.4). In §3 zal dan op basis van de verkregen inzichten uit deze literatuurstudie getracht worden om modellen te ontwerpen die, naast A, ook de andere invloedsfactoren op het fenotype in rekening te brengen, met als doel het fenotype dat dieren zullen 'realiseren' zo nauwkeurig mogelijk te voorspellen.

1.1 Fokwaardeschatting op basis van informatie over verwante dieren

1.1.1 Data

Bij fokwaardeschatting op basis van data van verwante dieren, kan onderscheid gemaakt worden in drie types data die kunnen verzameld worden van individuen uit de populatie:

1. Fenotypische 'prestaties'.
2. De afstamming van alle dieren in de populatie.
3. Gekende omgevingsfactoren met een invloed op het fenotype (bv. bedrijf waar het dier zich op bevindt en seizoen van afkalven).

De fenotypische prestaties en de omgevingsfactoren kunnen overzichtelijk worden weergegeven in vectoren/matrices, die onmiddelijk als input kunnen dienen voor bijvoorbeeld BLUP-fokwaardeschatting (§1.1.3), meestal worden hierbij de fenotypische waarnemingen relatief ten opzichte van het gemiddelde uitgedrukt door van alle waarnemingen het gemiddelde af te trekken. De afstammingsgegevens daarentegen kunnen niet als onmiddelijke input voor fokwaardeschattingen gebruikt worden en dienen eerst nog omgezet te worden in een matrix: de *pedigree relationshipmatrix* / additief genetische verwantschapsmatrix A (niet te verwarren met het additief genetisch effect in model (1.1)), waarbij het getal op positie (x,y) overeenkomt met $a_{X,Y}$,

de additief genetische verwantschapscoëfficiënt tussen individu X en Y. De additief genetische verwantschapscoëfficiënt wordt hierbij gedefinieerd door:

De additief genetische verwantschapscoëfficiënt $\alpha_{X,Y}$ tussen twee individuen X en Y is gelijk aan tweemaal de kans dat bij random allel-sampling voor een willekeurige locus, het gesamplede allel van individu X gelijk is aan het gesamplede allel van individu Y.

Voor het berekenen van de additief genetische verwantschapsmatrix A zijn algoritmes beschikbaar waarvoor verwezen wordt naar de wetenschappelijke literatuur. Indien verondersteld wordt dat het additief genetisch effect (met variantie σ_A^2) het resultaat is van een zeer groot (∞) aantal individuele allel-effecten, dan krijgt $\sigma_A^2 \alpha_{xy}$ een betekenis als de covariantie tussen het additief genetisch effect/de fokwaarde (BV) van individu X en het additief genetisch effect/de fokwaarde (BV) van individu Y. $\sigma_A^2 A$ is bijgevolg de variantie-covariantie matrix van de fokwaarden van de individuen in de populatie en is daardoor essentieel bij het schatten van fokwaarden volgens de BLUP-methodologie (§1.1.3).

1.1.2 Fokwaardebepaling: selectie-index theorie

Er bestaan 2 methodologiën om fokwaarden te schatten: selectie-index theorie en BLUP methodologie. Van deze twee methodologiën is selectie-index theorie (1.2) de meest eenvoudige. Deze vertoont sterke gelijkenissen met lineaire regressie doordat de fokwaarde van een individu (BV) wordt berekend op basis van een lineaire combinatie van de fenotypische waarnemingen (x_i) op n verwanten (incl. eventueel zichzelf). De coëfficiënten b_i in vergelijking (1.2) dienen in tegenstelling tot lineaire regressie echter niet bepaald te worden op basis van een trainingsdataset, maar kunnen op basis van theoretische argumenten afgeleid worden. Aangezien het afleiden van deze coëfficiënten buiten de scope van deze masterproef ligt, wordt hiervoor verwezen naar gespecialiseerde literatuur.

$$BV = \sum_{i=1}^{n} x_i b_i \tag{1.2}$$

Een groot nadeel van de selectie-index, is dat voor iedere nieuwe structuur van de dataset met informatie over verwante individuen (meer/minder verwante dieren, andere verwantschappen), de coëfficiënten b_i in (1.2) opnieuw afgeleid moeten worden, wat deze methode omslachtig maakt in settings met grote variabiliteit in databeschikbaarheid van fenotypische informatie met betrekking tot verwante individuen.

1.1.3 Fokwaardebepaling: BLUP

In tegenstelling tot fokwaardebepaling op basis van de selectie-index theorie, kan de BLUP-methodologie goed omgaan met grote variatie in het aantal verwante individuen waarover informatie beschikbaar is. De rekenkracht die vereist is voor het berekenen van BLUP-fokwaarden is echter wel hoger dan voor fokwaardebepaling via selectie-index theorie doordat vaak grote matrices vermenigvuldigd of geïnverteerd moeten worden. Hierdoor was de interesse voor BLUP-fokwaardebepaling ten tijde van de ontwikkeling van de methodologie door Henderson (1973) relatief beperkt, maar heeft deze met het ter beschikking komen van voldoende krachtige computers een hoge vlucht genomen. De BLUP-methodologie is gebaseerd op linear mixed models en werd initieel door Henderson (1973, 1984) ontwikkeld met als doel fokwaarden van dieren te schatten. Later werden linear mixed models ook in veel andere wetenschappelijke domeinen gebruikt. Het acroniem BLUP staat voor:

- *Best*: van alle mogelijke lineaire unbiased estimators/predictors, levert de BLUP-methodologie de kleinste MSE (Mean Squared Error) op.
- *Linear*: het model dat wordt gebruikt is lineair in zijn parameters.
- *Unbiased*: de verwachtingswaarde van de fokwaardeschatters is gelijk aan de werkelijke fokwaarden: $E[BV] = BV$
- *Prediction*: de uitdrukkingen die worden gebruikt om de fokwaarden te berekenen werden door Henderson predictors genoemd (om het verschil met Maximum Likelihood Estimators duidelijk te maken).

Om de outputvariabele $\boldsymbol{y}$ (het fenotype) te verklaren maakte Henderson gebruik van het linear mixed model:

$$\boldsymbol{y} = X\boldsymbol{\beta} + Z\boldsymbol{u} + \boldsymbol{\epsilon} \tag{1.3}$$

Met $\boldsymbol{y}$ een vector met de fenotypes, X de incidentie matrix[2] van de fixed effecten, $\boldsymbol{\beta}$ de vector met de coëfficiënten van de fixed effecten, Z de incidentie matrix van de random effecten, $\boldsymbol{u}$ de vector met de random effecten en $\boldsymbol{\epsilon}$ de de vector met fouttermen. Omdat model (1.3) vaak te veel vrijheidsgraden heeft om unieke parameterschatters te kunnen berekenen, legde Henderson een distributieve restrictie op aan de random effecten:

$$\boldsymbol{u} \sim N\left(\mathbf{0}, \sigma^2 G\right) \tag{1.4}$$

Na het opleggen van deze distributieve restrictie, kunnen er meestal wel een unieke parameterschatters $\hat{\boldsymbol{\beta}}$ en $\hat{\boldsymbol{u}}$ berekend worden. Dit kan gebeuren door de probabili-

[2]Een incidentie matrix geeft de relaties tussen twee 'klassen' weer aan de hand van een matrix bestaande uit nullen en enen, bijvoorbeeld tussen de individuen (rijen) en het geboortejaar (kolommen)

teitsdensiteit van de *joint* distributie van $\boldsymbol{y}$ en $\boldsymbol{u}$ (1.5) te maximaliseren naar $\boldsymbol{\beta}$ en $\boldsymbol{u}$, wat aanleiding geeft tot de mixed model equations van Henderson (1.6). Aangezien het onderscheid tussen fixed en random effecten geen fysische grondslag heeft en enkel maar wordt bepaald door op welke effecten er een distributieve restrictie wordt toegepast, is het niet altijd even duidelijk of een bepaald effect dan wel bij de fixed effecten of bij de random effecten moet ingedeeld worden. Zo kan de invloed van het geboortejaar op het fenotype ($\boldsymbol{y}$) vaak zowel als een fixed effect als als een random effect in model (1.3) ingebracht worden.

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix} \sim N\left(\begin{bmatrix} X\boldsymbol{\beta} \\ \boldsymbol{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} R & 0 \\ 0 & G \end{bmatrix} \right) \tag{1.5}$$

$$\begin{bmatrix} X^\top R^{-1} X & X^\top R^{-1} Z \\ Z^\top R^{-1} X & Z^\top R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} X^\top R^{-1} \boldsymbol{y} \\ Z^\top R^{-1} \boldsymbol{y} \end{bmatrix} \tag{1.6}$$

Voor het gebruik van model (1.3) voor (eenvoudige) fokwaardeschatting op dierniveau (animal-BLUP), kunnen de mixed model equations (1.6) onder bepaalde condities vaak vereenvoudigd worden. Indien bijvoorbeeld verondersteld wordt dat alle fenotypische 'prestaties' met dezelfde nauwkeurigheid werden geregistreerd en alle omgevingsfactoren die correlaties kunnen veroorzaken tussen de waargenomen fenotypes (bedrijf, geboortejaar,...) opgenomen zijn als fixed effecten in het model, dan is $\sigma^2 R$ gelijk aan $\sigma_E^2 I_n$, met σ_E^2 de residuele (milieu) variantie. Indien daarnaast enkel maar de additief genetische effecten (de fokwaarden) als random effect worden opgenomen, dan komt $\sigma^2 G$ overeen met $\sigma_A^2 A$, met σ_A^2 de variantie op de additief genetische effecten en A de additieve verwantschapsmatrix. Worden deze vereenvoudigingen doorgevoerd in (1.6), dan resulteert dit in (1.7), waarbij $\hat{\boldsymbol{u}}$ alle fokwaarden zal bevatten.

$$\begin{bmatrix} X^\top X & X^\top Z \\ Z^\top X & Z^\top Z + \frac{\sigma_E^2}{\sigma_A^2} A^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} X^\top \boldsymbol{y} \\ Z^\top \boldsymbol{y} \end{bmatrix} \tag{1.7}$$

1.2 Fokwaardeschatting op basis van genetische informatie

1.2.1 Merkers

Merkers die gebruikt worden voor de voorspelling van fokwaarden zijn meestal SNP's (*single nucleotide polymorphisms*). Deze SNP's kunnen de causale mutaties zijn die

verantwoordelijk zijn voor een bepaald fenotype, maar dat is niet noodzakelijk zo, SNP's in introns kunnen even goed functioneren als merkers.

Het voorspellen van fokwaarden op basis van SNP's is gebaseerd op linkage oneven-wicht. Dit principe houdt in dat, als de merkers voldoende dicht op elkaar liggen in het genoom, genen (met een effect op bijvoorbeeld de melkproductie) in linkage-onevenwicht zijn met nabijgelegen merkers, waardoor de aan- of afwezigheid van deze merkers kan gebruikt worden om te 'voorspellen' welk allel van een gen op een bepaalde locus zal aanwezig zijn. In theorie zou men dus op basis van de merker-informatie kunnen voorspellen welke allelen aanwezig zijn op iedere locus van een ieder gekend gen, om op basis hiervan vervolgens fokwaardevoorspellingen te doen. In de praktijk worden echter rechtstreeks op basis van de SNP-informatie fokwaar-den voorspeld, de onderliggende hypothese is nog altijd dat de aanwezigheid van een bepaalde SNP in de meerderheid van de gevallen ook gepaard gaat met de aan-wezigheid van bepaalde allelen op nabijgelegen loci, maar dit wordt niet expliciet opgenomen in de gebruikte modellen.

De genetische informatie die kan gehaald worden uit de resultaten van een SNP-analyse kan voor iedere bestudeerde SNP gecodeerd worden aan de hand van drie toestanden: afwezig (0), heterozygoot aanwezig (1) en homozygoot aanwezig (2). Op die manier kunnen de resultaten van een SNP-analyse voorgesteld worden in een vector, met als dimensie het aantal bestudeerde SNP's en als waarden de toestand waarin iedere SNP voorkomt (0,1,2).

1.2.2 Molecular breeding value

De fokwaardeschatting die gebeurt op basis van de SNP-informatie alleen noemt men de *molecular breeding value* (MBV) (Moser et al., 2009). Zoals eerder al besproken wordt hiervoor infomatie m.b.t. de aan- of afwezigheid van bepaalde SNP's (geco-deerd via 0, 1, 2) door een model als input gebruikt (Moser et al., 2009; Pintus et al., 2012), waarna het model een voorspelling doet over de fokwaarde van een bepaald kenmerk. De modellen worden getraind met gekende fokwaarden met een hoge be-trouwbaarheid (vaak bekomen via de klassieke BLUP-methodologie, §1.1.3), afkomstig van bijvoorbeeld stieren met enkele honderden nakomelingen. Na training kan het be-komen model dan gebruikt worden om de fokwaarden van jonge dieren (die nog geen nakomelingen hebben) te voorspellen. In het verleden is al een grote verscheidenheid aan modelstructuren uitgetest, gaande van de klassieke multiple lineaire regressie, over BLUP modellen en support vector regressie tot Bayesiaanse regressie. In wat volgt worden een aantal van deze modelstructuren besproken, elk met hun voor-en nadelen. Voor een bespreking van Bayesiaanse methoden wordt naar gespeciali-

seerde literatuur verwezen (vb. Meuwissen et al., 2001) omdat (1) het conceptuele framework van deze methoden sterk verschilt van het conceptuele framework van de meeste andere methoden en (2) de toepassing van deze methoden vaak zeer veel rekentijd vraagt (Moser et al., 2009), waardoor de toepassingsmogelijkheden binnen deze masterproef beperkt zijn.

Multiple lineaire regressie

Multiple lineaire regressie is een basistechniek die een zeer breed scala aan toepassingen heeft. Deze techniek kan worden voorgesteld door (1.8), met $\boldsymbol{y}$ een $(n \times 1)$ vector met geobserveerde waarnemingen, X een $(n \times p)$ designmatrix, $\boldsymbol{\beta}$ een $(p \times 1)$ parametervector en $\boldsymbol{\epsilon}$ een $(n \times 1)$ vector met fouttermen.

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{met} \quad \boldsymbol{\epsilon} \sim N\left(\mathbf{0}, \sigma^2 I_n\right) \tag{1.8}$$

Om de parameters in (1.8) te schatten, wordt de *quadratic loss* verliesfunctie (1.9) geminimaliseerd naar $\boldsymbol{\beta}$, wat resulteert in parameterschatter (1.10).

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta})^2 = \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 \tag{1.9}$$

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, L(\boldsymbol{\beta}) = (X^\top X)^{-1} X^\top \boldsymbol{y} \tag{1.10}$$

In het kader van MBV-schatting, kan (1.8) herschreven worden als (1.11), waarbij de index k alle merkers overloopt die in het model werden opgenomen ($k \leq m$, met m het aantal merkers), $x_{ik} \in \{0, 1, 2\}$ het aantal kopieën van SNP k is dat voorkomt in het genoom en β_k de *least squares* regressiecoëfficiënt is horende bij het additief genetisch effect van SNP k. Het model heeft geen term voor het intercept, aangezien de gemiddelde fokwaarde van de individuen in de populatie per definitie 0 is.

$$\boldsymbol{y} = \sum_{k} \boldsymbol{x_k} \beta_k + \boldsymbol{\epsilon} \tag{1.11}$$

Er zijn twee redenen waarom de index k meestal niet alle SNP's overloopt waar data over beschikbaar zijn. Een eerste reden is dat het aantal SNP's vaak het aantal dieren overtreft waarvan merkerdata beschikbaar zijn en waarvan de fokwaarden een voldoende hoge betrouwbaarheid hebben, waardoor het vaak onmogelijk is om unieke parameterschatters te vinden. Een tweede bezwaar, dat niet met de praktische beschikbaarheid van de data heeft te maken, is het probleem van multicollineariteit. Deze multicollineariteit wordt veroorzaakt doordat de merkers niet alleen sterk gecorreleerd zijn met de aan- of afwezigheid van bepaalde allelen op nabijgelegen *quantitative trait loci* (QTL), maar ook met de aan- of afwezigheid van nabijgelegen mer-

kers. Dit resulteert in het feit dat de schattingen van de regressiecoëfficiënten β_k van sterk gecorreleerde merkers onderhevig zijn aan grote standaardafwijkingen, wat negatieve effecten heeft op de predictieve kwaliteiten van het bekomen model. Multicollineariteit kan niet vermeden worden door een uitbreiding van de dataset, wat maakt dat, zelfs al zou het aantal dieren waarvoor merkerinformatie beschikbaar is de hoeveelheid merkers met een factor 10 of meer overschrijden, het niet mogelijk is om een multiple lineaire regressiemodel op te stellen dat alle beschikbare merkerinformatie meeneemt en goede predictieve capaciteiten heeft. De enige oplossing om problemen veroorzaakt door multicollineariteit te vermijden bij multiple lineaire regressie, is door van iedere 'set' van sterk gecorreleerde merkereffecten slechts één merkereffect op te nemen in het model. Door deze noodzakkelijke modelvereenvoudiging is multiple lineaire regressie echter niet in staat is om alle informatie te gebruiken die kan gehaald worden uit de merkers, wat meteen ook het grootste nadeel van deze techniek is.

Bij het selecteren van merkers die zullen worden opgenomen in het model gaat men vaak volgens een stapsgewijze procedure te werk, waarbij gestart word met een eenvoudig model (vb. $\boldsymbol{y} = \mathbf{0}$), waarna in iedere stap merkereffecten worden toegevoegd aan of verwijderd uit het model van de vorige stap, meestal op basis van een p-waarde significantiethreshold (eventueel bepaald via cross-validatie). Het finale model wordt bekomen als de procedure op een punt komt waarbij er geen merkers meer gevonden worden om aan het model toe te voegen of uit het model weg te laten (Moser et al., 2009).

Ridge en lasso regressie

Ridge en lasso regressie zijn twee technieken die grote gelijkenissen tonen met multiple lineaire regressie, maar wel in staat zijn om modellen met meer parameters dan waarnemingen te fitten. Daarnaast zijn ridge en lasso regressie ook in staat om de potentiële grote variatie op de parameterschattingen door multicollineariteit tussen de merkers (zie multiple lineaire regressie), voor een groot stuk te 'dempen', wat de predictieve kwaliteit van het bekomen model ten goede komt. Om dit te bereiken wordt de klassieke *quadratic loss* verliesfunctie (1.9) in beide technieken uitgebreid met een *shrinkage-penalty*, die er voor zorgt dat de parameters (behalve een eventueel intercept β_0) geregulariseerd worden naar nul toe en daardoor minder extreme waarden zullen aannemen in vergelijking met de regressiecoëfficiënten die men zou bekomen bij multiple lineaire regressie in geval van multicollineariteit. Voor een algemeen lineair model (1.8), wordt de verliesfunctie bij ridge regressie gegeven door (1.12) en bij lasso regressie door (1.13).

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{1.12}$$

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n}(y_i - \boldsymbol{x_i}^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \mid \beta_j \mid \tag{1.13}$$

In (1.12) is te zien dat bij ridge regressie de regularisatie van de parameterwaarden wordt bereikt door de kwadraten van de parameterwaarden (met uitzondering van β_0) op te nemen in de verliesfunctie, wat er voor zorgt dat de parameters die geen effect hebben een waarde dicht bij nul zullen aannemen. Bij lasso regressie wordt de regularisatie echter gerealiseerd door de absolute waarde van de parameters op te nemen in de verliesfunctie, wat er toe leidt dat parameters die geen effect hebben op de te modelleren variabele exact tot nul zullen worden herleid. Hierdoor heeft lasso-regressie ook toepassingen in het kader van modelselectie. Doordat de gemiddelde fokwaarde van een populatie per definitie gelijk is aan nul, is er voor het bepalen van molecular breeding values geen intercept nodig, waardoor (1.12) en (1.13) kunnen vereenvoudigd worden tot (1.14) en (1.15).

$$L(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \tag{1.14}$$

$$L(\boldsymbol{\beta}) = \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum \mid \beta_j \mid \tag{1.15}$$

Ridge regressie en lasso regressie werden door verschillende auteurs toegepast om MBV's te schatten, waarbij zowel ridge als lasso regressie modellen opleveren met (relatief) goede predictieve kwaliteiten (Usai et al., 2009; Ogutu et al., 2012; Li en Sillanpää, 2012; Piepho, 2009).

Principale componenten regressie en partial least squares regressie

Principale componenten analyse en partial least squares zijn technieken die werden ontwikkeld om p-dimensionale datasets (aantal variabelen = p) zodanig te transformeren dat de variabiliteit in de dataset zo veel mogelijk wordt geconcentreerd in de eerste dimensies van de getransformeerde dataset (met dimensie p). Waar bij principale componenten analyse deze 'condensatie' van de variabiliteit in de dataset het enige doel is, wordt daarnaast bij partial least squares ook getracht om de correlatie tussen de eerste dimensies van de getransformeerde dataset en de respons, zo hoog mogelijk te maken. Aangezien kennis over hoe de bijhorende transformatie-algoritmes te werk gaan, weinig bijdraagt tot inzicht in hoe principale componenten analyse en partial least squares kunnen gebruikt worden bij regressieproblemen, wordt in deze masterproef niet dieper op deze algoritmes ingegaan.

Eenmaal de p-dimensionale dataset getransformeerd is, kan de dimensionaliteit van de dataset gereduceerd worden door enkel maar de eerste m variabelen, ook wel componenten genoemd, van de getransformeerde dataset te weerhouden. Nadat op deze manier de dimensionaliteit van de dataset werd gereduceerd, kan multiple lineaire regressie toegepast worden op deze eerste m componenten, zonder dat overfitting of multicollineariteit een negatieve invloed hebben op de modelperformantie. De veronderstelling die hierbij wordt gemaakt, is dat de eerste m componenten niet enkel de meerderheid van de variabiliteit in de dataset bevatten, maar ook de meerderheid van de informatie die nuttig is in de context van het beschouwde regressieprobleem. Meestal volstaan hierbij waarden voor m (eventueel bepaald via cross-validatie) die een heel stuk kleiner zijn dan p om een model op te stellen met aanvaardbare predictieve kwaliteiten (Solberg et al., 2009; Moser et al., 2009; Pintus et al., 2012; Colombani et al., 2012).

BLUP

De BLUP methodiek voor het bepalen van de parameters van een linear mixed model beschreven in §1.1.3 kan ook gebruikt worden in het kader van predictie van molecular breeding values (Moser et al., 2009; Heslot et al., 2012; Pintus et al., 2012; Luan et al., 2009; Meuwissen et al., 2001; Ogutu et al., 2011). In het kader van BLUP-voorspelling van MBV, bevat het linear mixed model enkel een random effect voor de merkers, waardoor (1.3) kan vereenvoudigd worden tot:

$$\boldsymbol{y} = Z\boldsymbol{u} + \boldsymbol{\epsilon} \tag{1.16}$$

met $\boldsymbol{y}$ de fokwaarden die gebruikt worden om het model te trainen. Daarnaast wordt er meestal de veronderstelling gemaakt dat de random effecten van de SNP's onafhankelijk zijn van elkaar en verdeeld zijn volgens een normale distributie met gemiddelde 0 en variantie $\sigma_M^2 I$. Indien bovendien de betrouwbaarheid van de fokwaarden gebruikt om het model te trainen voldoende hoog is, kan verondersteld worden dat $\sigma^2 R \approx \sigma_E^2 I_n$, met σ_E^2 de residuele variantie op de gebruikte fokwaarden. Op basis van deze modelstructuur en de bijhorende veronderstellingen, kunnen de mixed model equations van Henderson (vgl. 1.6) vereenvoudigd worden tot (1.17), met $\lambda = \sigma_E^2/\sigma_M^2$.

$$\hat{\boldsymbol{u}} = (Z^\top Z + \lambda I)^{-1} Z^\top \boldsymbol{y} \tag{1.17}$$

Hierbij kan λ niet rechtstreeks bepaald worden op basis van de data en dient daarom een optimale waarde gekozen te worden op basis van bijvoorbeeld cross-validatie.

Support vector regressie

Support vector regressie (SVR) is een krachtige machine learning methode die in staat is om, met een beperkte rekenkracht, regressiemodellen op te stellen in hoog-dimensionale regressieruimtes. Vooraleer het regressieprobleem dat hierbij wordt gebruikt, uiteengezet kan worden, moeten echter eerst twee belangrijke concepten worden geïntroduceerd.

1. **kernels**
   Een klassieke methode om bij multiple lineaire regressie de modelperformantie van een model te verhogen, is het opnemen van extra variabelen in de model-structuur. Een voorbeeld hiervan in het geval van een 2-dimensionale waarnemingsruimte (x_1, x_2) is het introduceren van twee kwadratische effecten (x_1^2 en x_2^2) en een interactie-effect $(x_1 x_2)$ in het regressiemodel. Wiskundig geformuleerd wordt hierbij de tweedimensionale waarnemingsruimte geprojecteerd op een 5-dimensionale regressieruimte:

$$(x_1, x_2) \mapsto (x_1, x_2, x_1^2, x_2^2, x_1 x_2) \tag{1.18}$$

   Een nadeel van deze methode is dat bij hoog-dimensionale regressieruimten het expliciet projecteren van alle trainingsvectoren naar de regressieruimte en het berekenen van de regressiecoëfficiënten vaak veel rekenvermogen vraagt. Indien bijvoorbeeld over 100 000 merkers informatie beschikbaar is, dan heeft de waarnemingsruimte dimensie 100 000 en de regressieruimte die alle lineaire, kwadratische en 2e orde interactietermen omvat dimensie $100000 + 100000^2$. De hoge dimensie van de regressieruimte maakt het expliciet projecteren van alle waarnemingen op deze regressieruimte en het berekenen van alle regressiecoëfficiënten in deze regressieruimte computationeel zeer intensief. Er kan echter aangetoond worden dat voor bepaalde regressieruimten de bijhorende regressievergelijking kan geschreven worden als:

$$\hat{y}_i = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x, x_i) \tag{1.19}$$

   Met n het aantal waarnemingen en $K(x, x_i)$ een kernel. Concreet betekent dit dat de oplossing in bepaalde hoog-dimensionale regressieruimten kan herleid worden tot een oplossing met slechts $n + 1$ parameters. Om de parameterschattingen te bepalen, moet hierbij enkel voor ieder mogelijk paar van twee trainingsvectoren de kernelfunctie geëvalueerd worden, wat aanleiding geeft tot een $(n \times n)$ kernel matrix. Dit is computationeel meestal veel minder intensief dan de regressieruimte expliciet te construeren. Door het gebruik van kernels

kan dus een regressie worden uitgevoerd in een hoog-dimensionale regressieruimte, zonder dat deze expliciet dient geconstrueerd te worden. Er is een grote verscheidenheid aan kernels, de meest eenvoudige is de lineaire kernel (1.20) en twee van de meest populaire zijn de polynomiale kernel van graad d (1.21) en de Gausiaanse kernel / radiale kernel (1.22), met parameter γ.

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j} \tag{1.20}$$

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij} x_{i'j}\right)^d \tag{1.21}$$

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2\right) \tag{1.22}$$

2. ***epsilon-insensitive error* functie**

   Bij multiple lineaire regressie, ridge regressie en lasso regressie wordt telkens gebruik gemaakt van de *quadratic error* functie (1.23) om de afwijkingen van de modelvoorspellingen ten opzichte van de werkelijke responswaarden in rekening te brengen in de respectievelijke verliesfucties (1.9), (1.12) en (1.13).

$$V(y_i, \hat{y}_i) = (y_i - \hat{y}_i)^2 \tag{1.23}$$

   Bij support vector regressie wordt echter niet gebruik gemaakt van de *quadratic error* functie, maar van de *epsilon-insensitive error* functie, welke wordt gegeven door (1.24), met ϵ een vrij te kiezen parameter.

$$V(y_i, \hat{y}_i) = \begin{cases} 0 & \text{als } |y_i - \hat{y}_i| \leq \epsilon \\ |y_i - \hat{y}_i| - \epsilon & \text{anders} \end{cases} \tag{1.24}$$

   Wordt de *quadratic error* functie in eender welke verliesfunctie vervangen door de *epsilon insensitive error* functie, dan leidt dit er toe dat trainingsvectoren waarvoor de waarde van de respons minder dan ϵ afwijkt van de modelvoorspelling, geen bijdrage meer zullen leveren aan de verliesfunctie. Een gevolg hiervan is dat de partiële afgeleide van de verliesfunctie naar deze trainingsvectoren gelijk wordt aan nul, waardoor een (beperkte) verschuiving van deze goed voorspelde trainingsvectoren geen invloed zal hebben op de optimale parameterwaarden die zullen bekomen worden na minimalisatie van de verliesfunctie. Hierdoor kan het algoritme dat gebruikt wordt om de verliesfunctie te minimaliseren zich, bij wijze van spreken, 'concentreren' op de trainingsvectoren die slecht door het model worden voorspeld. De exacte parameterwaarden worden

dus volledig vastgelegd door de trainingsvectoren waarvoor de modelvoorspelling meer dan ϵ afwijkt van de werkelijke respons.

Kernels en de *epsilon insensitive error* functie vormen de basis van regressieprobleem (1.25) - (1.28), dat gebruikt wordt om een SVR op te stellen. Het gebruik van kernels in model (1.25) zorgt er hierbij voor dat er gebruik kan gemaakt worden van hoogdimensionale regressieruimten zonder dat deze expliciet moeten worden uitgerekend, terwijl de integratie van de *epsilon insensitive error* functie in verliesfunctie (1.26) er voor zorgt dat het algoritme dat de SVR opstelt, zich kan 'concentreren' op de moeilijk te voorspellen trainingsvectoren. Aangezien model (1.25), $n+1$ parameters bevat, met n het aantal waarnemingen, is het nodig om naast een *error penalty* ook een *complexity penalty* op te nemen in verliesfunctie (1.26), anders kunnen immers geen unieke parameterwaarden bepaald worden. Bij SVR wordt deze *complexity penalty* geïmplementeerd door een gewogen som van de kwadraten van de parameters, met uitzondering van het intercept, mee te nemen in de verliesfunctie. De wegingsfactor λ is hierbij een hyperparameter, waarvoor via cross-validatie een optimale waarde kan worden gekozen.

$$\hat{y}_i = \beta_0 + \sum_{j=1}^{n} \alpha_j K(x_i, x_j) \tag{1.25}$$

$$L(\beta_0, \alpha_1, \ldots, \alpha_n) = \sum_{i=1}^{n} V(y_i, \hat{y}_i) + \lambda \sum_{i=1}^{n} \alpha_i^2 \tag{1.26}$$

$$V(y_i, \hat{y}_i) = \begin{cases} 0 & \text{als } |y_i - \hat{y}_i| \leq \epsilon \\ |y_i - \hat{y}_i| - \epsilon & \text{anders} \end{cases} \tag{1.27}$$

$$\{\hat{\beta}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_n\} = \underset{\beta_0, \alpha_1, \ldots, \alpha_n}{\operatorname{argmin}} L(\beta_0, \alpha_1, \ldots, \alpha_n) \tag{1.28}$$

Verschillende auteurs hebben SVR toegepast voor het schatten van MBV's. Uit deze studies blijkt dat bij gebruik van een radiale kernel SVR meestal beter presteert dan BLUP (pg. 10) (Moser et al., 2009; Honarvar en Ghiasi, 2013; Long et al., 2011), terwijl dit bij gebruik van lineaire kernels niet het geval is (Ogutu et al., 2011; Heslot et al., 2012; Long et al., 2011).

Neurale netwerken

Neurale netwerken (NN) zijn een klasse van modellen die zowel kunnen gebruikt worden voor classificatieproblemen als voor regressieproblemen. Een van de grote sterktes van deze neurale netwerken is dat ze in staat zijn om een grote verscheidenheid aan functies te benaderen, zonder dat dit grote wijzigingen in de architectuur van het model vereist. Dit maakt neurale netwerken zeer geschikt voor regressie- of clas-

sificatieproblemen waarbij er weinig kennis is omtrent de functionele vorm van het verband tussen de gemeten inputs en de waargenomen waarden (bv. bij beeldverwerking). Bij het schatten van additief genetische effecten in het kader van MBV, is de functionele vorm van het verband tussen merkerdata en de fokwaarden echter bij benadering gekend, waardoor neurale netwerken zelden een meerwaarde bieden in vergelijking met eenvoudigere machine-learning technieken (Heslot et al., 2012; Okut et al., 2013; Sinecen, 2019).

In Figuur 1.1 wordt de structuur van een eenvoudig neuraal netwerk weergegeven. Het neuraal netwerk in Figuur 1.1 omvat 1 *input layer* $(x_1, ..., x_p)$, 1 *hidden layer* $(z_1, ..., z_M)$ en 1 *output layer* $(y_1, ..., y_K)$. Bij het modelleren van (ongekende) outputs op basis van waargenomen inputs, krijgt iedere knoop een waarde toegewezen. De *input layer* krijgt, zoals de naam al doet vermoeden, de waarden van de inputparameters toegewezen, waarbij één knoop wordt voorzien per inputparameter. Analoog krijgt de *output layer* de outputwaarden van het model toegewezen, bij regressieproblemen met één knoop in de outputlayer ($K = 1$) is dit bijvoorbeeld de voorspelde outputwaarde, maar bij classificatieproblemen met meerdere klassen (bv. cijferherkenning op foto's), kunnen er meerdere outputknopen aanwezig zijn, die elk de kans uitdrukken dat de inputvector behoort tot de klasse die werd toegewezen aan de outputknoop (in het geval van cijferherkenning zijn deze klassen $0, 1, ..., 9$, wat resulteert in 10 outputknopen). De waarde die een knoop toegewezen krijgt, hangt enerzijds af van de waarden van de knopen in de onderliggende laag waarmee de bestudeerde knoop is verbonden en anderzijds de activatiefunctie van de knoop. In de meeste neurale netwerken wordt deze activatiefunctie gegeven door een niet-lineaire functie (bv. een sigmoïdale functie) die wordt toegepast op een lineaire functie van de waarden van de knopen in de onderliggende laag, wat zichtbaar is in de wiskundige beschrijving van het neuraal netwerk in Figuur 1.1 in vergelijkingen (1.29) - (1.33), waar $f_m()$ en $g_k()$ deze niet-lineaire functies representeren.

$$\boldsymbol{x}^\top = \begin{bmatrix} x_1 & x_2 & \cdots & x_P \end{bmatrix} \tag{1.29}$$

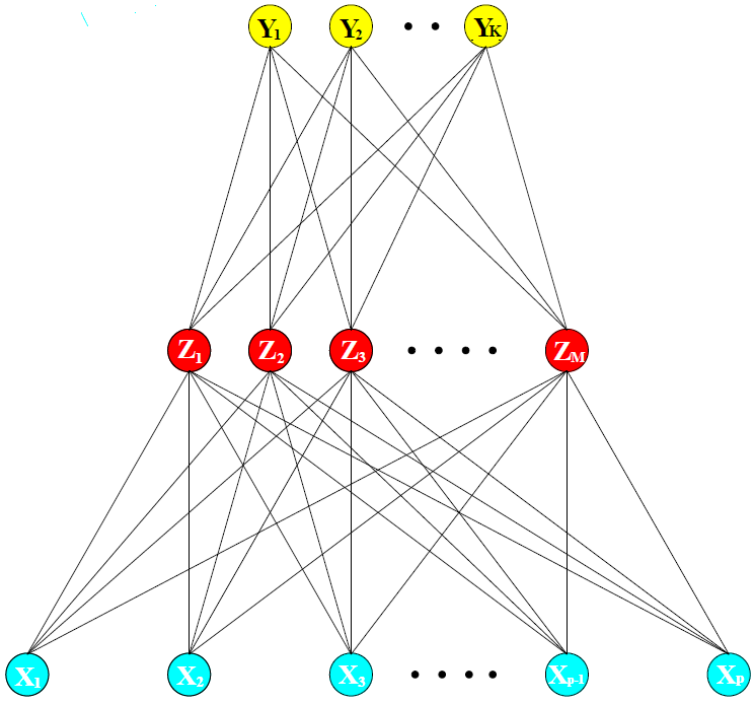
$$z_m = f_m\left(\alpha_{0m} + \boldsymbol{x}^\top \boldsymbol{\alpha_m}\right),\ m = 1, \ldots, M \tag{1.30}$$

$$\boldsymbol{z}^\top = \begin{bmatrix} z_1 & z_2 & \cdots & z_m & \cdots & z_M \end{bmatrix} \tag{1.31}$$

$$y_k = g_k\left(\beta_{0k} + \boldsymbol{z}^\top \boldsymbol{\beta_k}\right),\ k = 1, \ldots, K \tag{1.32}$$

$$\boldsymbol{y}^\top = \begin{bmatrix} y_1 & y_2 & \cdots & y_k & \cdots & y_K \end{bmatrix} \tag{1.33}$$

Om het neuraal netwerk in Figuur 1.1 te 'trainen' op basis van een dataset met n waarnemingen, kan gebruik gemaakt worden van een *gradient descent* methode om



Figuur 1.1: Opbouw van een neuraal netwerk in zijn meest eenvoudige vorm (Hastie et al., 2008)

verliesfunctie (1.34) te minimaliseren. Deze methode wordt in het kader van neurale netwerken echter meestal *back-propagation* genoemd, ondanks dat de gebruikte methode niet verschilt van de *gradient-descent* methode die bij veel andere optimalisatieproblemen wordt gebruikt. Net zoals bij ridge- en lassoregressie, bestaat ook bij neurale netwerken de verliesfunctie (1.34) uit enerzijds een *error penalty* en anderzijds een *complexity penalty* om overfitting te vermijden. Indien voor een neuraal netwerk alle parameterwaarden worden samengebracht in een vector $\boldsymbol{\theta}$, dan kan de *back-propagation* update van de parameters in iteratie $(r+1)$ voorgesteld worden door (1.35), met γ een constante die ook wel de '*learning rate*' wordt genoemd.

$$L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \|\boldsymbol{y_i} - \hat{\boldsymbol{y}}_{\boldsymbol{i}}\|^2 + \lambda \left(\sum_{m=1}^{M} \|\boldsymbol{\alpha_m}\|^2 + \sum_{k=1}^{K} \|\boldsymbol{\beta_k}\|^2 \right) \tag{1.34}$$

$$\boldsymbol{\theta}^{(r+1)} = \boldsymbol{\theta}^{(r)} - \gamma \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{(r)}} \tag{1.35}$$

1.2.3 Genomic breeding value

Stacking

In §1.1 werd besproken hoe een fokwaarde van een dier kan berekend worden op basis van informatie van verwante dieren (PBV, *pedigree breeding value*) en in §1.2.2 werden verschillende modellen besproken die gebruikt kunnen worden om op basis van merkerdata *molecular breeding values* (MBV) te berekenen. Geen enkele van al deze modellen gebruikt echter zowel data van verwante dieren als merkerinformatie om fokwaarden te berekenen. Door geschatte PBV's echter te combineren met geschatte MBV's, kunnen *genomic breeding values* (GBV) bekomen worden, welke potentieel een hogere betrouwbaarheid hebben dan PBV of MBV op zichzelf. De meest eenvoudige manier om een PBV met een MBV te combineren, is door het gemiddelde van beiden te nemen. Een uitbreiding hiervan is om een gewogen gemiddelde te nemen van PBV en MBV, waarbij de gewichten kunnen bepaald worden op basis van de betrouwbaarheid (R^2) van beide fokwaardeschattingen (Moser et al., 2009):

$$GBV = \frac{w_1 MBV + w_2 PBV}{w_1 + w_2} \tag{1.36}$$

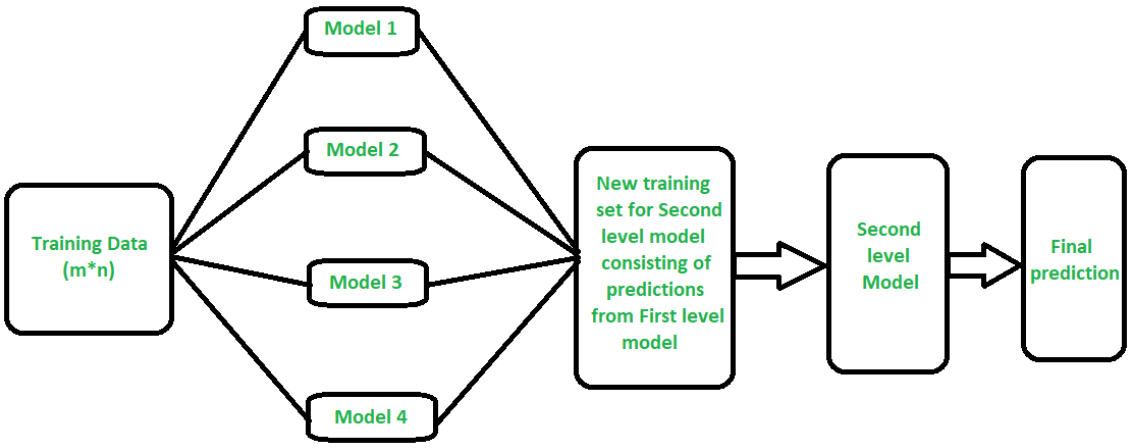
met

$$w_i = \frac{R_i^2}{1 - R_i^2} \tag{1.37}$$

Deze methode is intuïtief correct, maar niet optimaal. Zo is het eenvoudig in te zien dat indien men een PBV verkregen via een BLUP-animal model combineert met 3 verschillende MBV's (verkregen via bv. ridge regressie, support vector regressie en een neuraal netwerk) via een vergelijking analoog aan (1.36), dat de invloed van de merkerdata niet in evenwicht is met de invloed van de info over verwante dieren.

Een oplossing voor dit probleem is *stacking*, een methodologie waarbij de finale modelvoorspelling (hier de GBV) bepaald wordt in twee fasen (Figuur 1.2). In de eerst fase worden er op basis van de inputvariabelen met verschillende modellen/modelstructuren outputvariabelen voorspeld, waarna in de tweede fase alle voorspellingen uit de eerste fase (eventueel aangevuld met de oorspronkelijke inputs) gebruikt worden als inputs voor een samenvattend model dat de finale voorspelling maakt. Toegepast op de voorspelling van GBV, zou dit er op neer komen dat in een eerste fase via verschillende modelstructuren PBV's en MBV's zouden worden voorspeld, waarna al deze PBV's en MBV's (eventueel aangevuld met hun respectieve betrouwbaarheden en bepaalde delen van de initiële informatie) als inputs worden doorgegeven aan een

samenvattend model dat al deze data en voorspellingen combineert om tot een finale voorspelling van de GBV te komen.



Figuur 1.2: Schematisch overzicht van stacking (Geeks for Geeks, 2019)

ss-GBLUP

Alle eerder besproken methodes kunnen voor het berekenen van GBV's enkel maar gebruikt worden in een tweestapsprocedure, waarbij in een eerste stap de MBV en PBV apart worden berekend en in een tweede stap, via bv. *stacking*, deze aparte fokwaarden geïntegreerd worden in de finale GBV-voorspelling. Er bestaat echter een variant op het klassieke pedigree-BLUP model (1.7) die toelaat om in één stap GBV-fokwaardeschattingen te maken van alle dieren in de populatie: single step Genomic Best Linear Unbiased Prediction of kortweg ssGBLUP. In deze methode wordt model (1.7) gebruikt, maar de pedigree-verwantschapsmatrix A wordt vervangen door een verbeterde versie (H), die opgesteld wordt door zowel gebruik te maken van merker-informatie als van pedigree-informatie.

Om de berekening van H overzichtelijk te kunnen weergeven, dienen A en $\boldsymbol{u}$ in (1.7) eerst geherstructureerd te worden, zodat de fokwaarden van de l individuen zonder merkerinformatie zich in de eerste l posities van $\boldsymbol{u}$ bevinden en de fokwaarden van de k individuen met merkerinformatie zich in de laatste k posities van $\boldsymbol{u}$ bevinden. Dit laat toe om zowel $\boldsymbol{u}$ als A in een blokmatrixnotatie weer te geven (1.38), waarbij de submatrices van A een eenduidige consistentie hebben[3].

$$\boldsymbol{u} = \begin{bmatrix} \boldsymbol{u_1^{(l)}} \\ \boldsymbol{u_2^{(k)}} \end{bmatrix} \qquad A = \begin{bmatrix} A_{11}^{(l\times l)} & A_{12}^{(l\times k)} \\ A_{21}^{(k\times l)} & A_{22}^{(k\times k)} \end{bmatrix} \tag{1.38}$$

[3] A_{11}: verwantschappen tussen niet-merkergeteste individuen
A_{22}: verwantschappen tussen merkergeteste individuen
$A_{12}\, A_{21}$: verwantschappen tussen wel en niet merkergeteste individuen

Voor de berekening van H wordt naast de pedigree-informatie die vervat zit in A ook gebruik gemaakt van de genomische verwantschappen tussen de k individuen met merkerinformatie, welke worden weergegeven in de $(k \times k)$ genomic-verwantschapsmatrix G. Om G te berekenen, wordt er vertrokken van de gecodeerde (zie §1.2.1) merkerinformatie die per individu beschikbaar is (1.39), samen met de allelfrequenties van de m merkers in de populatie (1.40).

$$\boldsymbol{x}^\top = \begin{bmatrix} x_1 & x_2 & \cdots & x_i & \cdots & x_m \end{bmatrix} \quad x_i \in \{0, 1, 2\} \tag{1.39}$$

$$\boldsymbol{f}^\top = \begin{bmatrix} f_1 & f_2 & \cdots & f_i & \cdots & f_m \end{bmatrix} \quad f_i \in [0, 1] \tag{1.40}$$

Na centrering van de k vectoren $\boldsymbol{x}$ via (1.41) zodat $E[\boldsymbol{w}] = \boldsymbol{0}$, worden de k verkregen vectoren $\boldsymbol{w}$ samengevoegd tot een $(k \times m)$ matrix W (1.42), welke in (1.43) samen met $\boldsymbol{f}$ wordt gebruikt om de genomic-verwantschapsmatrix G te berekenen (Van Raden, 2008; Druet et al., 2014).

$$\boldsymbol{w} = \boldsymbol{x} - 2\boldsymbol{f} \tag{1.41}$$

$$W^\top = \begin{bmatrix} \boldsymbol{w_1} & \boldsymbol{w_2} & \cdots & \boldsymbol{w_i} & \cdots & \boldsymbol{w_k} \end{bmatrix} \tag{1.42}$$

$$G = \frac{WW^\top}{2\boldsymbol{f}^\top(\boldsymbol{1} - \boldsymbol{f})} \tag{1.43}$$

Naast het gebruik van G op zichzelf voor verdere berekeningen, is het ook mogelijk om verder te werken met een gewogen gemiddelde van G en A_{22} (1.44) (Christensen en Lund, 2010), wat toelaat om een ss(G)BLUP-model op te stellen dat het midden houdt tussen de klassieke pedigree-BLUP en de zuivere ssGBLUP.

$$G^* = \lambda G + (1 - \lambda) A_{22} \tag{1.44}$$

De genomic-pedigree-verwantschapsmatrix H kan vervolgens worden berekend via vergelijking (1.45), welke werd afgeleid op basis van principes uit de probabiliteitstheorie (Legarra et al., 2009). Er kan daarnaast aangetoond worden dat H^{-1} kan berekend worden via (1.46)(Aguilar et al., 2010; Christensen en Lund, 2010), welke veel eenvoudiger is dan (1.45) en bovendien intuïtief logisch is opgebouwd. Voor het berekenen van GBV's via ssBLUP, dient A^{-1} in (1.7) enkel maar vervangen te worden door (1.46), waarna na oplossen van de mixed model equations de GBV's worden bekomen.

$$H = \left[\begin{array}{c:c} A_{11} - A_{12}A_{22}^{-1}A_{21} + A_{12}A_{22}^{-1}GA_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}G \\ \hdashline GA_{22}^{-1}A_{21} & G \end{array}\right] \tag{1.45}$$

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix} \tag{1.46}$$

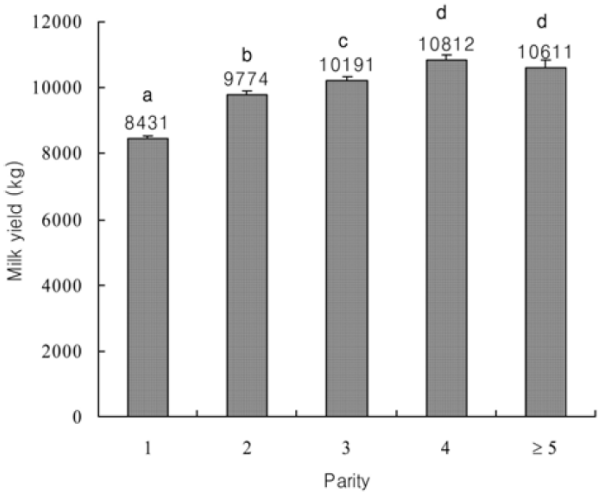
De integratie van merkerinformatie zorgt er voor dat de genomic-pedigree-verwantschapsmatrix H de werkelijke verwantschappen beter weergeeft dan de pedigree-verwantschapsmatrix A, wat er toe leidt dat de betrouwbaarheid van de GBV's bekomen via ssGBLUP groter is dan de betrouwbaarheid van fokwaarden berekend via de klassieke pedigree-BLUP. Daarnaast laat ssGBLUP ook toe om de fokwaarden van wel- en niet-merkergeteste individuen met één model uit te rekenen, waarbij maximaal gebruik kan gemaakt worden van de beschikbare fenotypische informatie. Hierdoor is ssGBLUP niet alleen een veelgebruikte methode in de wetenschappelijke literatuur (Aguilar et al., 2010; Colombani et al., 2012; Honarvar en Ghiasi, 2013; Sinecen, 2019), maar wordt ssGBLUP ook voor praktijktoepassingen gebruikt, bijvoorbeeld bij de berekening van genomic fokwaarden bij het VPF (Vlaamse Piétrain Fokkerij).

1.3 Invloed van niet-genetische factoren op de melkproductie

1.3.1 Pariteit

Het productiepotentieel van koeien stijgt tot en met de 4e lactatie (Lee en Kim, 2006) (Figuur 1.3). De reden hiervan is dat als een vaars op een leeftijd van twee jaar afkalft, dit geen volgroeide koe is, waardoor ze niet dezelfde voederinname (Oldenbroek, 1989) en mobilisatie van lichaamsreserves (Lee en Kim, 2006) kan realiseren als een koe met een hogere pariteit. Daarnaast heeft een vaars nog een deel van de energie die ze opneemt nodig voor haar eigen groei, wat er voor zorgt dat het aandeel van de opgenomen energie dat bij een vaars beschikbaar is voor de melkproductie lager is dan bij een koe met een hoger pariteit.

Het moment waarop de piekproductie bij vaarzen wordt bereikt is niet significant verschillend van het moment waarop koeien hun piekproductie bereiken (Hansen et al., 2006), maar doordat koeien hun lactatie starten met een hogere melkgift dan vaarzen en bovendien nog eens sneller stijgen in productie dan vaarzen, is de piekproductie van koeien hoger dan van vaarzen (Friggens et al., 1999; Hansen et al., 2006). De melkproductie bij vaarzen is echter wel persistenter dan bij koeien (Schutz et al., 1990; Friggens et al., 1999), waardoor het verschil in piekproductie tussen koeien en vaarzen proportioneel groter is dan het uiteindelijke verschil in 305-d productie.



Figuur 1.3: Gemiddelde 305-d productie (kg) van koeien in functie van de pariteit, verschillende letters geven significante verschillen aan (Lee en Kim, 2006)

1.3.2 Temperatuur en luchtvochtigheid

Door hun hoge metabolische activiteit zijn melkkoeien gevoelig voor hittestress: omstandigheden waarbij koeien de warmte die vrijkomt bij hun metabolisme moeilijker aan de omgeving kunnen afgeven. Hierbij kunnen temperatuur en luchtvochtigheid niet los van elkaar gezien worden, aangezien beiden een invloed hebben op hoe makkelijk een koe warmte kan overdragen naar haar omgeving. Zo maakt een hogere omgevingstemperatuur warmteafgifte via convectie en conductie moeilijker, terwijl een hogere relatieve luchtvochtigheid evaporatieve warmteafgifte bemoeilijkt. In het kader van hittestress worden deze twee parameters daarom vaak geïntegreerd in één hittestress-indicator: de temperatuur-luchtvochtigheidsindex (THI, *temperature humidity index*), welke op basis van de droge-boltemperatuur (T, °C) en de relatieve luchtvochtigheid (RH, relative humidity, 0 - 100%) kan berekend worden via formule (1.47) (National Research Council, 1971). Hittestress begint vanaf een THI van 72, eens $THI > 78$ spreekt men van ernstige hittestress en als $THI > 89$ spreekt men van zeer ernstige hittestress.

$$THI = (1.8T + 32) - (1 - RH/100) * (T - 14.4) \tag{1.47}$$

Naast formule (1.47), zijn er ook formules beschikbaar voor het berekenen van de THI waarin de relatieve luchtvochtigheid vervangen wordt door de natte-boltemperatuur of de dauwpuntstemperatuur van de lucht (Dikmen en Hansen, 2009).

Hittestress leidt tot een verlaagde voederopname, een verlaagde melkproductie, lagere vet- en eiwitgehaltes in de melk en een verminderde vruchtbaarheid (Dunn et al., 2014; Davis et al., 2003; Bryant et al., 2007; Bouraoui et al., 2002). De lagere melkproductie wordt hierbij veroorzaakt door het feit dat een koe met hittestress haar voederopname reduceert, zodat haar metabolische warmteproductie verlaagt (Bouraoui et al., 2002; Bryant et al., 2007). Doordat de koe echter minder voeder opneemt, zijn er ook minder nutriënten beschikbaar voor de melkproductie, waardoor deze daalt (Bryant et al., 2007). Het primair effect van hittestress is dus een verlaagde voederopname, met als secundair effect een verlaagde melkproductie (West et al., 2003). Doordat het effect van een lagere voederopname op de melkproductie een zekere lag-tijd kent, vinden veel onderzoeken sterke verbanden en correlaties tussen de THI van de vorige 1-3 dagen en de melkproductie (West et al., 2003; Bouraoui et al., 2002; Bryant et al., 2007). THI-waarden van opeenvolgende dagen zijn echter vaak sterk gecorreleerd, waardoor er ook verbanden kunnen gevonden worden tussen de melkproductie en de THI van de dag zelf, maar deze verbanden zijn vaak minder sterk (Bouraoui et al., 2002; West et al., 2003). De grootte van de daling in melkproductie varieert in functie van de ernst van de hittestress (Tabel 1.1) en kan oplopen tot meer dan 30% productieverlies ten opzichte van de normale melkproductie (Bouraoui et al., 2002).

Tabel 1.1: Productiedaling bij Holsteinkoeien ten gevolge van hittestress. In kolom 'THI' wordt via codering de gebruikte THI-waarde weergegeven, vb. (0,-1,-2): het gemiddelde van de THI op de metingsdag zelf en de twee dagen ervoor

bron	**THI**	**kritische THI**	**productieverlies (THI^{-1})**	
West et al. (2003)[a]	(-2)		kg M:	-0.88
West et al. (2003)[a]	(0)		kg M:	-0.69
Bouraoui et al. (2002)	(0)	69	kg M:	-0.41
Bryant et al. (2007)	(0,-1,-2)	68	g V+E:	-10

[a] In de studie van West et al. (2003) werd geen kritische waarde voor THI bepaald

De gevoeligheid voor hittestress is deels genetisch bepaald, zo blijkt uit verschillende proeven dat Holsteinkoeien gevoeliger zijn voor hittestress dan Jerseykoeien (Bryant et al., 2007; West et al., 2003) en konden Ravagnolo en Misztal (2000) aantonen dat binnen een ras een deel van de gevoeligheid voor hittestress verklaard kan worden via additief genetische effecten, wat betekent dat er selectie mogelijk is naar runderen die meer hitte-resistent zijn. Ravagnolo en Misztal (2000) rapporteren hierbij daarnaast ook dat de genetische correlatie tussen melkproductie en hittestress-tolerantie -0.3 is, waardoor de sterke selectie op productiviteit in het verleden er voor heeft

gezorgd dat moderne koeien gevoeliger zijn voor hittestress dan hun soortgenoten van enkele decennia geleden. De correlatie is echter beperkt in grootteorde, waardoor Ravagnolo en Misztal (2000) aangeven dat selectie op beide kenmerken tegelijk mogelijk moet zijn.

1.3.3 Jongveeopfok - leeftijd bij eerste kalving

De leeftijd bij eerste kalving (LEK) heeft een grote invloed op de productiekengetallen die een vaars zal realiseren tijdens haar eerste lactatie, zo vonden Nor et al. (2013) bij een studie uitgevoerd in 2010 op 100 Nederlandse melkveebedrijven (Tabel 1.2) dat vaarzen met een LEK van 22 maanden een 305d-productie realiseerden die 432 kg M (≈ 1.4 kg M/d) lager lag dan de 305d-productie van vaarzen die 4 maanden later, op een leeftijd van 26 maanden, voor het eerst kalfden. Indien alle afkalfleeftijden relatief werden uitgedrukt tegenover de bedrijfsmediaan, werden analoge resultaten bekomen: binnen eenzelfde bedrijf hebben dieren die 2 maanden vroeger dan de bedrijfsmediaan afkalven gemiddeld een 337 kg M lagere 305-dagenproductie (≈ 1.1 kg M/d) in de eerste lactatie dan dieren die 4 maanden ouder zijn (+2 in Tabel 1.2) als ze voor het eerst afkalven.

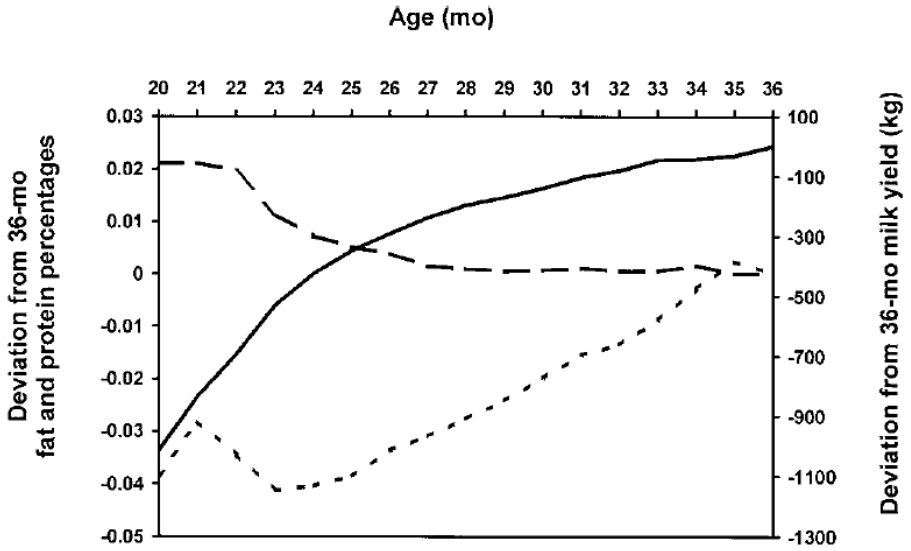
Tabel 1.2: Effect van leeftijd bij eerste kalving op de 305-d melkproductie (Nor et al., 2013). In de kolom Δ worden de verschillen in 305d-melkproductie weergegeven ten opzichte van de referentieleeftijd, welke aangeduid is met *

absolute LEK			relatieve LEK		
LEK (mnd)	**305d-prod. (kg M)**	**Δ (kg M)**	**LEK [a] (mnd)**	**305d-prod. (kg M)**	**Δ (kg M)**
≤ 20	6295	-869	≤ -6	6274	-998
21	6697	-467	-5	7125	-147
22	6876	-288	-4	6738	-534
23	7021	-143	-3	6952	-320
24*	7164	0	-2	7098	-174
25	7212	48	-1	7182	-90
26	7308	144	0*	7272	0
27	7454	290	+1	7358	86
28	7489	325	+2	7435	163
29	7545	381	+3	7542	270
30	7540	376	+4	7591	319
31	7760	596	+5	7643	371
≥ 32	7816	652	≥ +6	7779	507

[a] Relatief ten opzichte van de bedrijfsmediaan

Gelijkaardige resultaten werden bekomen door Pirlo et al. (2000), die 1 048 942 eerste lactaties van Italiaanse vaarzen analyseerde (Figuur 1.4). Net zoals in Tabel 1.2, is er in Figuur 1.4 een duidelijke trend zichtbaar die toont dat een kleinere LEK gepaard gaat met een lagere 305-d productie. De daling van de 305-d productie tegenover

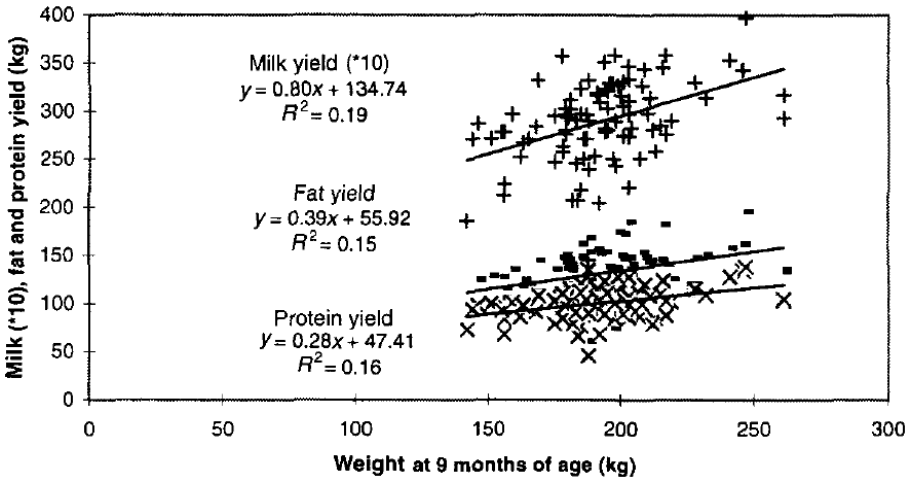
vaarzen met een LEK van 36 maanden is hierbij relatief beperkt (< 400 kg M) zolang de LEK groter is dan ±24 maanden. Indien de LEK kleiner wordt, dan treedt er een sterke daling op van de 350-d productie die de vaarzen in hun eerste lactatie gemiddeld zullen realiseren (tot −1000 kg melk bij een LEK van 20 maanden). Wat betreft het vetgehalte van de melk is op Figuur 1.4 te zien dat dit, net als de 305-d productie, een stijgende trend vertoont bij een stijgende LEK van 23 naar 36 maanden, terwijl de LEK nagenoeg geen invloed heeft op het eiwitpercentage van de melk bij een LEK tussen 23 en 36 maanden. Het grillige patroon voor de curves van het vet- en eiwitpercentage tussen 20 en 23 maanden is te wijten aan een beperkt aantal waarnemingen met een hoge variabiliteit en is daarom niet relevant (Pirlo et al., 2000).



Figuur 1.4: Effect van leeftijd bij eerste kalving op de 305-d melkproductie (volle lijn), het vetpercentage (stippellijn) en het eiwitpercentage(streepjeslijn), de effecten worden relatief uitgedrukt tegenover de prestaties van vaarzen die op een leeftijd van 36 maanden voor het eerst afkalven (Pirlo et al., 2000)

Ook Berry en Cromie (2009) bekwamen analoge resultaten na analyse van 196 120 eerste lactaties van Ierse vaarzen die tussen 2000 en 2006 voor het eerst kalfden: naarmate de LEK stijgt, stijgt ook de 305-d productie die in de eerste lactatie wordt gerealiseerd. De consistente waarneming over verschillende bronnen heen dat een hogere LEK resulteert in een hogere 305d-productie, betekent echter niet dat het economisch verantwoord is te streven naar een hogere LEK, aangezien een hogere LEK ook gepaard gaat met een langere opfokperiode, wat tot een lagere productie per levensdag leidt als vaarzen met een hoge LEK op het einde van de eerste lactatie worden vergeleken met vaarzen met een lage LEK (Lin et al., 1986).

De daling van de 305-d productie in eerste lactatie bij een kleinere LEK is volgens Hietanen en Ojala (1995) voornamelijk te wijten aan een lager lichaamsgewicht van vaarzen die op een jongere leeftijd voor het eerst kalven. Hierbij is het volgens Macdonald et al. (2005) niet het lagere lichaamsgewicht zelf dat de directe oorzaak is van de verlaagde 305d-productie, maar wel de verlaagde DS-opnamecapaciteit die hiermee gepaard gaat. Deze positieve correlatie tussen levend gewicht en melk-, vet- en eiwitproductie in eerste lactatie werd bevestigd door Van der Waaij et al. (1997). Hij woog kalveren/pinken op een leeftijd van 9 en 21 maanden en bekwam dat 1 kg extra levend gewicht op een leeftijd van 9 (21) maanden leeftijd resulteerde in 8.03 (6.65) kg meer melk, 0.39 (0.25) kg meer vet en 0.28 (0.19) kg meer eiwit tijdens de eerste lactatie (Figuur 1.5). Dit bevestigt het belang van een goede jongeveeopfok indien men om economische redenen enerzijds streeft naar een lage LEK en anderzijds toch hoge producties tijdens de eerste lactatie beoogt.



Figuur 1.5: Invloed van het gewicht op 9 maanden leeftijd op de melk-, vet- en eiwitproductie in de eerste lactatie (Van der Waaij et al., 1997)

1.3.4 Melkregime

Het aantal keer dat een koe per dag wordt gemolken heeft een invloed op de hoeveelheid melk, vet en eiwit die ze produceert. In België en Nederland worden koeien doorgaans 2 keer per dag gemolken, maar er zijn ook andere melkregimes in gebruik, waarbij koeien vaker (3 maal per dag) of minder vaak (1 maal per dag) gemolken worden.

Koeien die 3 maal per dag gemolken worden, produceren in vergelijking met koeien die tweemaal per dag worden gemolken gemiddeld 15% (10-17) meer melk, 10% (5-13) meer vet en 10% (7-14) meer eiwit. De stijging in de geproduceerde hoeveelheid

vet en eiwit is lager dan voor de hoeveelheid melk omdat de vet- en eiwitpercentages in de melk bij driemaal daags melken iets dalen (Klei et al., 1997; Campos et al., 1994; Smith et al., 2002).

In meer extensieve, vaak grasgebaseerde, melkveehouderijsystemen wordt soms vanwege strategische/economische redenen overgegaan op een melkregime waarbij de koeien maar 1 keer per dag worden gemolken (Stelwagen et al., 2013). Meestal wordt hierbij maar tijdens het laatste deel van de lactatie overgegaan op éénmaal daags melken. Indien men koeien echter gedurende de hele lactatie slechts éénmaal melkt, leidt dit tot gemiddeld 32% (30-35) minder melk, 29% (25-31) minder vet en 29% (26-32) minder eiwit, waarbij het vet- en eiwitgehalte in de melk iets hoger is dan bij tweemaal daags melken (Holmes et al., 1992; Clark et al., 2006; Rémond et al., 2004). Daarnaast heeft eenmaal daags melken ook een negatief effect op de persistentie van de melkgift (Stelwagen et al., 2013).

Uit studies waarbij verschillende uierkwartieren van een koe met een verschillend melkregime werden gemolken, blijkt dat de verhoogde/verlaagde melkproductie bij respectievelijk drie- of eenmaal per dag melken niet systemisch, maar lokaal wordt gereguleerd (Bernier-Dodier et al., 2010; Stelwagen en Knight, 1997; Sorensen et al., 2001; Wall en McFadden, 2007). Op korte termijn, bijvoorbeeld vlak na het verlagen van de melkfrequentie van 2 melkbeurten per dag naar 1 melkbeurt per dag, spelen vooral fysiologische effecten een rol bij de daling van de melkproductie. Zo zorgt een hogere druk van geproduceerde melk in de uier (± 17-18 h na de vorige melkbeurt (Stelwagen en Knight, 1997)) er voor dat de *tight junctions* beginnen lekken en ondervinden uierparenchymcellen bij lange melkintervallen meer mechanische stress, wat via mechanotransductie kan worden omgezet in intracellulaire signalen die de melkproductie remmen (Stelwagen et al., 1995).

Op middellange en lange termijn wordt de daling/stijging van de melkproductie vooral door veranderingen in de genexpressie van het uierparenchym veroorzaakt. Zo is er bijvoorbeeld bij koeien die eenmaal daags gemolken worden meer apoptose (geprogrameerde celdood) van secreterende cellen onder invloed van een hogere expressie van apoptose-genen in cellen van het uierparenchym (Littlejohn et al., 2009; Grala et al., 2011) en is de expressie van genen voor de productie van vet, eiwit en lactose lager in vergelijking met koeien die twee maal per dag worden gemolken (Grala et al., 2011).

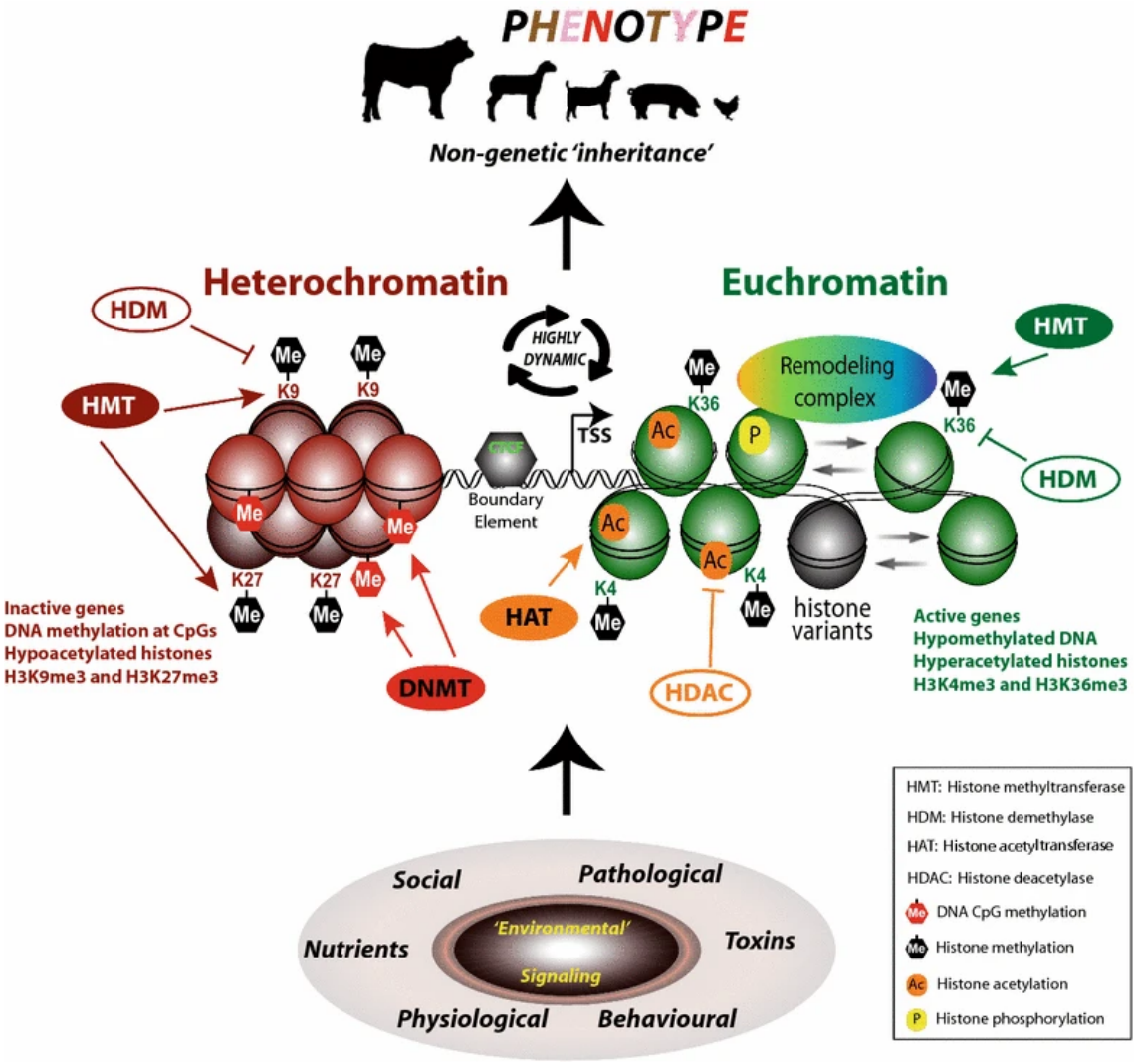
1.4 Invloed van epigenetische factoren op het fenotype

Epigenetica verwijst naar stabiele wijzigingen in genexpressie die niet gepaard gaan met wijzigingen in de DNA-sequentie (Feil, 2006; Scholtz et al., 2014). De term stabiel kan hierbij zowel verwijzen naar stabiliteit gedurende opeenvolgende celdelingen (Scholtz et al., 2014; Singh et al., 2012), als stabiliteit gedurende opeenvolgende generaties (González-Recio, 2012; Singh et al., 2012). In de ruime betekenis omvat epigenetica dus zowel genetisch geregelde wijzigingen in genexpressie die plaatsvinden tijdens bijvoorbeeld celdifferentiatie (tot bv. adipocyt, neuron,..) (Scholtz et al., 2014; Singh et al., 2012), als milieu-geïnduceerde wijzigingen in genexpressie (met een mogelijke invloed op het fenotype) (Scholtz et al., 2014). Hierdoor is epigenetica moeilijk te plaatsen in model (1.1), aangezien epigenetica enerzijds elementen bevat die duidelijk tot *G* behoren, anderzijds elementen bevat die duidelijk tot *E* behoren en daarbovenop nog elementen bevat die moeilijk in één van deze twee categorieën zijn in te delen.

Het geheel van alle moleculaire elementen die aanwezig zijn in de nabijheid van het DNA en betrokken zijn bij het regelen van de genexpressie, wordt het epigenoom genoemd. Er zijn verschillende moleculaire mechanismen gekend die hierbij een rol spelen, waarbij het niet uitgesloten is dat er naast de reeds gekende mechanismen nog andere bestaan. De meest bestudeerde worden hieronder kort besproken.

- **DNA-methylatie**
  Bij DNA-methylatie wordt de genexpressie van bepaalde genen onderdrukt door methylgroepen op de DNA-streng aan te brengen (Figuur 1.6). Dit gebeurt hoofdzakelijk op CpG dinucleotiden (5'- cytosine - guanine - 3'), waarbij cytosine wordt omgezet in 5-methylcytosine doordat enzymes (DNA-methyltransferases, DNMT) op de 5' positie van de pyrimidinering in cytosine een methylgroep plaatsen (Triantaphyllopoulos et al., 2016).

- **Post translationele modificatie van histonen**
  In eukaryote cellen komen de chromosomen onder normale omstandigheden niet voor als een vrije DNA-streng, maar als een chromatinedraad bestaande uit nucleosomen. Deze nucleosomen bestaan uit een octameer van histonen (2 × H3, 2 × H4, 2 × H2A en 2 × H2B) waarrond de DNA-streng is gewonden (Triantaphyllopoulos et al., 2016). De genexpressie van het DNA dat om het nucleosoom is gewonden, of zich in de omgeving van het nucleosoom bevindt, kan gewijzigd worden door modificatie van de histonen (acetylatie, fosforylatie, methylatie,



Figuur 1.6: Moleculaire verklaring voor de invloed van milieuomstandigheden op de genexpressie en het fenotype (Triantaphyllopoulos et al., 2016)

ubiquitylatie,...) (Figuur 1.6). Hierdoor kan de chromatinestructuur veranderen of kunnen er bindingsplaatsen ontstaan voor effectoren die de genexpressie op- of neerreguleren (Bannister en Kouzarides, 2011).

- **Chromatine remodelling**

  De densiteit van de chromatine heeft invloed op de toegankelijkheid van het DNA voor transcriptie en dus ook op de genexpressie. Hierbij is het zo dat DNA in dense chromatine (heterochromatine) veel minder toegankelijk is voor transcriptie dan DNA in regio's met minder nucleosomen per lengte-eenheid van de DNA-streng (euchromatine) (Ho en Crabtree, 2010). Het hermodelleren van de chromatine gebeurt door ATP-afhankelijke *chromatin remodelling complexes* (Figuur 1.6) die in staat zijn de densiteit van de chromatine te regelen door nucleosomen toe te voegen, te verwijderen of te herstructureren (Ho en Crabtree, 2010).

- **Niet-coderende RNAs (ncRNA)**
  Ook niet-coderende stukjes RNA (ncRNA) spelen een rol in de epigenetica, zo hebben ze bijvoorbeeld een belangrijke invloed op chromatine remodelling en histon modificatie (Triantaphyllopoulos et al., 2016). Hierbij kan een onderscheid gemaakt worden tussen epigenetische effecten waarbij de ncRNA's een initiërende/tijdelijke functie vervullen en effecten waarbij de permanente aanwezigheid van ncRNA's vereist is (Keller en Bühler, 2013). Beide types van ncRNA-epigenetica kunnen hierbij voorkomen in hetzelfde organisme, zo hebben ncRNA's bijvoorbeeld enkel maar een initiërende functie bij de vorming van heterochromatine op de *silent mating locus* van de gist *Schizosaccharomyces pombe* (Jia et al., 2004; Hall et al., 2002), terwijl voor de vorming en het behoud van centromerische heterochromatine in diezelfde gist continue aanwezigheid van het verantwoordelijke ncRNA is vereist (Volpe et al., 2002).

Op een hoger niveau onderscheidt Scholtz et al. (2014) op zijn beurt 3 mechanismen die betrokken zijn bij het tot stand komen van een bepaald epigenoom in een bepaald individu:

- **Parental imprinting**
  Parental imprinting is een proces dat enkel maar bij de zoogdieren voorkomt (m.u.v. de *Monotremata*) (Jirtle en Weidman, 2007). Hierbij wordt tijdens de gametogenese het epigenetisch patroon in bepaalde regio's van het genoom gewist en vervangen door een nieuw epigenetisch patroon (de imprint) dat afhankelijk is van het geslacht van het individu waarin de gametogenese plaatsvind (paternale vs. maternale imprint) (Jirtle en Weidman, 2007; Scholtz et al., 2014; Triantaphyllopoulos et al., 2016). Indien de paternale (maternale) imprint bestaat uit een verhoogde methylatie van het DNA, leidt dit in de nakomelingen doorgaans tot een (partiële) silencing van het *imprinted* paternale (maternale) allel, waardoor enkel maar het maternale (paternale) allel tot expressie komt in de nakomelingen (Neugebauer et al., 2010; Scholtz et al., 2014; Triantaphyllopoulos et al., 2016).

  Een gekend voorbeeld van parental imprinting is het callipyge allel bij schapen, dat musculaire hypertrofie veroorzaakt. Dit allel komt enkel maar tot expressie indien het heterozygoot aanwezig is (overdominantie) en bovendien afkomstig is van de vader (parental imprinting) (Cockett et al., 1996).

  Bij runderen zijn reeds 20 genen geïdentificeerd die gevoelig zijn aan imprinting (Triantaphyllopoulos et al., 2016) en heeft parentale imprinting invloed op minstens 10 kenmerken van runderkarkassen (Neugebauer et al., 2010).

- **Milieu-invloed op het epigenoom**
  Het milieu kan wijzigingen in het epigenoom (en dus ook in het fenotpe) induceren (Figuur 1.6). Hierbij is vooral in de prenatale en neonatale levensfase het epigenoom gevoelig voor milieu-invloeden. De term milieu dient hierbij heel breed geïnterpreteerd te worden en omvat naast klassieke parameters zoals bijvoorbeeld omgevingstemperatuur ook minder klassieke parameters zoals de metabolische toestand van de moeder tijdens de dracht en de sociale interacties tussen ouderdieren en hun nakomelingen.

  Het meest gekende voorbeeld van milieu-invloeden op het epigenoom heeft betrekking op de Nederlandse hongerwinter in 1944-1945: de kinderen van vrouwen die tijdens deze hongersnood zwanger waren, hadden niet enkel een lager geboortegewicht dan gemiddeld, maar waren op latere leeftijd ook gevoeliger voor allerlei gezondheidsproblemen (diabetes, obesitas, cardiovasculaire aandoeningen,...) (Lumey, 1992).

  Het is vooral rond dit type van epigenetische invloeden dat bij melkvee onderzoek is uitgevoerd, waarbij de meeste studies focussen op de invloed van de in-utero omstandigheden op de productiekengetallen die het embryo/de foetus later als vaars zal realiseren. Zo werden door González-Recio et al. (2012) vrouwelijke kalveren die werden geboren uit koeien (die lacteerden tijdens de dracht) vergeleken met vrouwelijke kalveren die werden geboren uit pinken. Bij de analyse van de data werden de fenotypische prestaties (melkproductie, vet/eiwitverhouding en productieve levensduur) hierbij eerst gecorrigeerd voor alle genetische en omgevingsfactoren die gekend waren: additief genetische effecten (de fokwaarde), LEK, jaar-regio effecten en jaar-kudde effecten. Zodat bijvoorbeeld het effect van een gemiddeld hoger genetisch potentieel van de kalveren geboren uit vaarzen in vergelijking met kalveren geboren uit koeien de resultaten van de statistische analyse niet kon beïnvloeden. Uit de resultaten, die samengevat worden in Tabel 1.3, besloot González-Recio et al. (2012) dat vaarskalveren die geboren werden uit koeien een significant ($p<0.02$) lagere 305d-melkproductie realiseerden als vaars in vergelijking met vaarskalveren die uit pinken werden geboren. Daarnaast is de productieve levensduur van vaarskalveren geboren uit koeien lager dan de productieve levensduur van vaarskalveren geboren uit pinken ($p<0.10$) en heeft de melk van vaarzen geboren uit koeien een hogere vet/eiwit verhouding ($p<0.001$), wat geassocieerd wordt met een hoger risico op metabole stoornissen (bv. ketose).

  Naast een invloed op productiekenmerken, kan de omgeving tijdens de neonatale levensfase ook een grote invloed hebben op het latere gedrag van een dier. Zo wordt bijvoorbeeld het moederinstinct bij ratten epigenetisch beïnvloed door

Tabel 1.3: Epigenetisch effect van lactatie tijdens de dracht op het fenotype van de nakomelingen. In de tabel worden de gemiddelde effecten weergegeven die werden bekomen na een Bayesiaanse analyse van de data, waarbij vaarskalveren geboren uit pinken als referentie werden gebruikt (González-Recio et al., 2012)

lactatienr.[a]	kg melk[b] (305d)	productieve levensduur (d)	vet/eiwit ratio (×100)
0	0	0	0
1	-18	-23	+0.49
2	-47	-15	+0.36
3+	-91	-9	+0.40

[a] lactatienr. van de moeder tijdens de dracht
[b] 305d-melkproductie (kg) tijdens de eerste lactatie

het gedrag van de moeder ten opzichte van haar jongen: likt een vrouwelijke rat frequenter dan de gemiddelde rat haar jongen (wat wordt beschouwd als een goede moedereigenschap), dan induceert dit wijzigingen in het epigenoom van haar jongen, waardoor haar dochters later ook betere moedereigenschappen zullen hebben en bovengemiddeld vaak hun jongen zullen likken. Eén van de epigenetische wijzigingen die hierbij plaatsvindt, is een verlaagde methylatie van de ER1b promotor van het ERα gen in het mediaal pre-optisch gebied van de hersenen. Dit gen codeert voor een oestrogeenreceptor die een rol speelt in de oestrogeen-geïnduceerde aanmaak van oxytocine receptors in de hersenen, waarvan geweten is dat ze een grote rol spelen bij het 'moederinstinct'. Het feit dat het in dit geval om epigenetische wijzigingen gaat die niet worden overgeërfd, maar worden geïnduceerd door het gedrag van de (pleeg)moeder, kon duidelijk aangetoond worden aan de hand van experimenten waarbij jongen werden verlegd van zorgzame moeders naar minder zorgzame moeders en omgekeerd (Young et al., 1998; Keverne en Curley, 2004; Francis et al., 1999; Champagne et al., 2001, 2006). Dit voorbeeld toont dat epigenetische informatie maternaal kan worden doorgegeven naar de volgende generatie, zonder dat deze epigenetische informatie moet bewaard blijven tijdens de meiose en de embryogenese. Een essentiële voorwaarde hierbij is wel dat er contact moet zijn tussen de moeder en haar jongen. Indien deze interactie onmogelijk wordt gemaakt, zoals vaak bij melkvee en nagenoeg altijd bij pluimvee het geval is, kan deze overdracht van epigenetische informatie immers niet meer plaatsvinden.

- **Doorgifte van epigenetische informatie via de gameten**
  Bij de embryogenese wordt een groot deel van het epigenetisch patroon gewist en vervangen door een epigenetisch patroon nodig voor de embryonale ontwikkeling. In een aantal regio's van het genoom gebeurt deze reset van het epigenoom echter niet of niet volledig, waardoor o.a. methylatiepatronen van het

DNA in deze genoomregio's kunnen overgedragen worden naar de volgende generatie. Dit maakt het mogelijk dat naast genetisch geregelde epigenetica (bv. voor celdifferentiatie) die vanzelfsprekend stabiel overgedragen kan worden via de gameten, ook bepaalde milieu-geïnduceerde epigenetische informatie gedurende een aantal generaties kan worden doorgegeven via de gameten.

Zo kon Braunschweig et al. (2012) bijvoorbeeld aantonen dat indien men grootouderberen (F0) een voeder geeft verreikt aan methyldonoren, dit in de F2 nog steeds aanleiding geeft tot significante ($p<0.05$) en net niet significante ($p<0.10$) verschillen in karkaseigenschappen in vergelijking met een controlegroep (o.a. percentage schouder van het karkas en parameters m.b.t. de vetheid van de dieren). Daarnaast kon Braunschweig et al. (2012) voor bepaalde genen ook significante verschillen in genexpressie aantonen tussen de F2 groep en een controlegroep.

HOOFDSTUK 2

PRIMAIRE DATAVERWERKING

In dit hoofdstuk wordt eerst de gebruikte soft- en hardware besproken, waarna de ruwe data worden beschreven en besproken wordt hoe vanuit deze ruwe data de inputvariabelen voor de machine-learning modellen werden opgesteld.

2.1 Soft- en hardwarespecificaties

Alle code die gebruikt werd tijdens de primaire dataverwerking en het trainen en evalueren van machine-learning modellen, werd geschreven in Python. Voor de visualisatie van de resultaten werd, naast Python, ook R gebruikt. Alle code werd opgesteld en gedebugd op een computer met een Intel Core i5 processor en 8 Gb RAM-geheugen. Voor het uitvoeren van code met een looptijd van meerdere uren, werd daarnaast ook gebruik gemaakt van de hpc-infrastructuur van de UGent.

2.2 Ruwe data

De ruwe data werden ter beschikking gesteld door Coöperatie CRV en werden aangeleverd met behulp van 7 tekstbestanden, die elk een deel van de informatie bevatten: afstamming, fokwaarden, aan- en afvoergegevens, afkalfgegevens, dagproducties, 305d-producties en reproductiedata. Bij het opstellen van de dataset (december 2019) werd een deel van de Nederlandse dieren waarvan CRV genomic fokwaarden beschikbaar heeft, geselecteerd en werd vervolgens alle mogelijks relevante informatie uit de CRV-databank geëxtraheerd. Een concrete implicatie hiervan is dat alle informatie die beschikbaar was op het moment dat de dataset werd samengesteld, ook werd gebruikt voor de fokwaardeberekeningen.

De afstammingsgegevens bevatten van 266 590 dieren het levensnummer, de geboortedatum, het geslacht, de rasbalk en het levensnummer van de moeder en de vader. Het oudste dier in deze deeldataset werd geboren in 1912, het jongste in 2019. Het levensnummer is een uniek identificatienummer dat in de hele dataset

wordt gebruikt om te verwijzen naar een bepaald dier. De rasbalk geeft weer wat de bijdrage is van verschillende rundveerassen aan het genoom van een bepaald dier, zo wordt de rasbalk van een dier met 3 raszuivere Holstein Friesian grootouders en één MRY grootouder genoteerd als: 3/4 HF 1/4 MRY. Bij de bepaling van de rasbalk werd een nauwkeurigheid gehanteerd van 1/8, wat inhoudt dat dieren waarvoor minder dan 12.5% van de genoom afkomstig is van een vreemd ras, als raszuiver worden beschouwd.

De fokwaarden voor 74 uiteenlopende kenmerken (Tabel 2.1) waren beschikbaar van 105 689 dieren. Hierbij werd door CRV zowel de genomic fokwaarde als de pedigree-gebaseerde fokwaarde aangeleverd. Voor een groot deel van de kenmerken was daarnaast ook de betrouwbaarheid van de fokwaarde aanwezig, dit zowel voor de genomic fokwaarde als voor de pedigree-gebaseerde fokwaarde. Zoals in Tabel 2.1 kan gezien worden, zijn de meeste courant gebruikte fokwaarden aanwezig in de dataset, met uitzondering van de fokwaarden % vet, % eiwit, % lactose en lichaamsgewicht, welke in de dataset ontbreken.

De aan- en afvoergegevens geven voor 137 961 dieren weer op welke datum het dier werd geboren/aangevoerd en op welke datum het dier werd afgevoerd. Iedere periode (van aanvoer/geboorte tot afvoer) wordt hierbij aan een specifiek bedrijf gelinkt, dat wordt geïdentificeerd aan de hand van een uniek bedrijfsnummer (UBN). 84 570 dieren volbrachten hun hele leven op het bedrijf van geboorte en hebben logischerwijs slechts één record. 53 391 dieren veranderden tijdens hun leven minstens één keer van bedrijf en hadden bijgevolg 2 of meer records. Het aantal keer dat een dier van bedrijf verhuisde varieerde hierbij tussen 1 en 26 keer.

De afkalfgegevens bevatten gegevens over 363 098 kalvingen, waarbij minimaal het levensnummer van de moeder en de kalfdatum zijn gegeven. Indien gekend, zijn daarnaast ook de drachtduur (aantal waarnemingen n = 217 664), het geboorteverloop (vlot, normaal, zwaar, keizersnede, andere hulp of afgezaagd; n = 236 983), het geslacht van de nakomeling (n = 236 705) en het geboortegewicht van het kalf (n = 198 836) opgenomen in de dataset.

De deeldataset die informatie bevat over de dagproducties, is de omvangrijkste deeldataset en bevat gegevens over 3 016 991 melkproductieregistraties (mpr's) van 99 598 unieke dieren, uitgevoerd tussen 1990 en 2019. Voor iedere mpr worden 12 variabelen bewaard: pariteit, kalfdatum, proefmelkdatum, de (on)geldigheid van de mpr, kg melk, % vet, % eiwit, % lactose, celgetal, ureumconcentratie, het aantal melkingen per dag en de status van de koe (normaal, driespeen, tochtig, uieronsteking, monstername onmogelijk, ziek, vers gekalfd, te vroeg gekalfd). Het aantal mpr's per

Tabel 2.1: Fokwaarden aanwezig in de dataset. In de tabel wordt voor iedere fokwaarde het gemiddelde (μ) en de standaardafwijking (σ) weergegeven. Afkortingen: afk. afkorting; int. interval; ins. inseminatie; glob. globuline; lv. levensvatbaarheid; opn. opname; voorsp. voorspellers

Fokwaarde	**Afk.**	μ	σ	**Fokwaarde**	**Afk.**	μ	σ
kg melk	kgM	-5	686	Karakter	KA	100	3.0
kg vet	kgV	3	22	Vleesindex	VLi	100	3.1
kg eiwit	kgE	0	18	Persistentie	PS	100	3.6
kg lactose	kgL	0	33	Laatrijpheid	LR	102	3.0
INET	INET	5	111	kg DS opn. lactatie 1	DM1	-0.3	0.7
NVI	NVI	7	65	kg DS opn. lactatie 2	DM2	-0.2	0.8
Directe levensduur	dld	71	177	kg DS opn. lactatie 3+	DM3	-0.2	0.9
Levensduur	lvd	70	180	kg DS opn.	dDM	-0.2	0.8
Better life gezondheid	BLh	-0.3	2.4	kg DS opn. met voorsp.	iDM	0.0	1.0
Better life efficiëntie	BLe	1.4	5.0	Frame	F	100	4.1
Uiergezondheid	Ugh	101	2.7	Type	R	101	3.1
Klinische mastitis	CM	101	2.2	Totaal uier	U	100	4.5
Subklinische mastitis	SCM	101	2.9	Totaal beenwerk	B	99	3.1
Vruchtbaarheid	Vru	100	3.1	Totaal exterieur	Ext	100	3.6
Non-return 56 dagen	N56	100	3.3	Hoogtemaat	HT	100	4.8
Int. afkalven-eerste ins.	IAI	101	3.4	Voorhand	VH	101	3.8
Tussenkalftijd	TKT	101	3.2	Inhoud	IH	100	4.2
Int. eerste-laatste ins.	IFL	100	2.7	Openheid	OH	100	3.7
Conception rate koe	CR	100	3.5	Conditiescore	CS	101	3.9
Conception rate pink	CRp	100	2.9	Kruisligging	KL	100	4.1
Leeftijd eerste ins. pink	AFI	100	3.5	Kruisbreedte	KB	99	4.2
Leeftijd eerste kalving	ALV	100	2.9	Beenstand achter	BA	100	2.9
Geboorte index	Gin	99	3.6	Beenstand zij	BZ	100	3.7
Geboortegemak	Geb	100	3.6	Klauwhoek	KH	99	3.4
Directe lv.	LVg	100	3.4	Beengebruik	BG	100	2.9
Afkalfgemak	Afk	100	4.3	Vooruieraanhechting	VA	100	4.3
Maternale lv.	LVa	100	3.2	Voorspeenplaatsing	VP	100	4.0
Klauwgezondheid	CLW	100	2.2	Speenlengte	SL	99	4.3
Ketose	KET	100	3.7	Uierdiepte	UD	100	4.6
Celgetal	Cgt	101	3.4	Achteruierhoogte	AH	100	4.3
Ureum	UR	0.2	1.7	Achterspeenplaatsing	AP	99	4.1
Kalvervitaliteit	SUR	100	3.1	Ophangband	OB	100	3.9
Draagtijd	DrD	101	3.6	Caseïne	CAS	-1	12
Geboortegewicht	Ggw	100	3.1	Beta-lactoglob.	BLG	2	37
Melksnelheid	MS	100	3.7	Alfa-lactalbumine	ALA	0	6
Melkrobot eficiëntie	EFF	99	3.2	Immunolactoglob. G	IgG	-1	31
Melkrobot interval	INT	101	3.1	Bovine serum A	BSA	-1	15

dier varieert van 1 tot 286 en het aantal mpr's per lactatie is minimaal 1 en maximaal 185.

Naast de dagproducties zelf, leverde CRV ook een deeldataset aan met (voorspelde) 305d-producties. Deze dataset bevat gegevens over 314 632 lactaties van 95 414 unieke dieren, waarbij volgende variabelen voor iedere lactatie aanwezig zijn: pariteit, kalfdatum, aantal lactatiedagen, einddatum van de lactatie (laatste mpr-datum indien nog ongekend), lactatiewaarde en de (voorspelde) 305d-productie, uitgedrukt in kg melk, kg vet en kg eiwit. De lactatiewaarde (lw) is hierbij een kengetal dat het economisch rendement van een lactatie tracht uit te drukken, relatief ten opzichte van het bedrijfsgemiddelde (CRV, 2015).

De laatste deeldataset bevat gegevens over reproductie-gerelateerde events: kunstmatige inseminatie, natuurlijke dekking, samenweiding met een stier, drachtcontrole, geregistreerde tochtigheden, embryo implantatie, gust verklaren, abortus en spoeling. In totaal bevat deze deeldataset informatie over 1 245 141 reproductie-gerelateerde handelingen bij 121 185 unieke dieren. Voor ieder event wordt steeds het levensnummer van het betreffende dier, de datum en het type event gegeven. Indien het een event betreft dat conceptie tot doel heeft, wordt daarnaast ook de gebruikte stier vermeld en indien het een drachtcontrole betreft, wordt de uitslag van de drachtcontrole vermeld (drachtig/niet drachtig).

2.3 Van ruwe data naar inputdata voor de modellen

De data die werden aangeleverd door CRV, zouden in theorie onmiddellijk kunnen gebruikt worden als inputdata voor de op te stellen machine learning modellen, dit door voor ieder dier alle informatie te verzamelen, een selectie uit te voeren op basis van het moment dat men wil bestuderen (bv. bij geboorte zijn er geen mpr-data beschikbaar van het dier zelf) en vervolgens al deze variabelen kolomsgewijs te concateneren tot een grote dataset die kan gebruikt worden om een machine-learning model op te trainen. Echter, de meeste machine-learning modellen behalen hogere performanties indien de inputdataset waarmee ze worden getraind eenduidig gedefinieerde variabelen bevat, wat bij het simpelweg kolomsgewijs concateneren van alle beschikbare informatie niet het geval is.

Het beste voorbeeld hiervan zijn de mpr-data. Zoals eerder vermeld, vertonen deze een grote variatie in het aantal mpr-metingen per lactatie: dit varieert tussen 1 en 185. Deze grote variatie wordt niet enkel veroorzaakt door het feit dat nog niet alle lactaties waarover informatie beschikbaar is, zijn afgerond (dan is het aantal mpr's beperkt), maar ook doordat de mpr-frequentie verschilt tussen de bedrijven. Voor be-

paalde koeien is om de 5 weken een mpr beschikbaar, voor andere koeien is dit om de 4 weken en voor nog andere koeien is om de paar dagen een mpr beschikbaar. Daarom leidt het simpelweg kolomsgewijs concateneren van alle mpr-gegevens niet tot een dataset met eenduidig gedefinieerde variabelen, zo is de tiende mpr bij bepaalde dieren op het einde van de eerste lactatiemaand beschikbaar, terwijl de tiende mpr voor andere dieren de laatste mpr van de lactatie is. Aangezien de informatie die kan afgeleid worden uit een mpr op het einde van de eerste lactatiemaand totaal verschillend is van de informatie die kan afgeleid worden uit een mpr op het einde van de lactatie, is het bijgevolg weinig waarschijnlijk dat kolomsgewijs concateneren van alle beschikbare data aanleiding zal geven tot modellen met een hoge performantie.

Daarom werden de ruwe data eerst verwerkt tot variabelen met een duidelijk afgelijnde betekenis, welke vervolgens gebruikt werden om datasets op te stellen om machine-learning modellen te trainen en te evalueren. Deze primaire dataverwerking wordt in chronologische volgorde besproken, zo is het eerst noodzakelijk om de mpr-data te verwerken (§2.3.2) vooraleer alle beschouwde productiekengetallen konden berekend worden (§2.3.8). De productiekengetallen zijn op hun beurt dan weer noodzakelijk om een interactie-effect tussen de bedrijven en de fokwaarden in rekening te kunnen brengen, zonder dat hierbij het aantal variabelen in de dataset in grote mate toeneemt (§2.3.9).

Ieder hierna besproken onderdeel van de primaire dataverwerking levert een subdataset op die telkens een deel van alle beschikbare informatie bevat. Bij het samenstellen van de datasets die werden gebruikt om de machine-learning modellen te trainen en te evalueren, werden de data uit deze verschillende subdatasets opgehaald en op een passende manier gecombineerd. Dit liet toe om op een snelle en efficiënte manier de meer dan duizend datasets op te stellen die nodig waren om alle modellen te trainen en evalueren, zonder dat hierbij telkens opnieuw vertrokken moest worden van de ruwe data.

2.3.1 Fokwaarden

De deeldataset met fokwaarden die door CRV werd aangeleverd, bevat zowel genomic fokwaarden als pedigree-gebaseerde fokwaarden, samen met hun respectievelijke betrouwbaarheid. De gemiddelde betrouwbaarheid van de genomic fokwaarden was beduidend groter dan die van de pedigree gebaseerde fokwaarden. Zo bedroeg de gemiddelde betrouwbaarheid voor het kenmerk kg melk 79.8% voor de genomic fokwaarden, terwijl deze maar 54.2% was voor de pedigree-gebaseerde fokwaarden. Analoog bedroeg voor het kenmerk levensduur de gemiddelde betrouwbaarheid van de genomic fokwaarde 62.7%, terwijl deze voor de pedigree-gebaseerde fokwaarde

maar 47.1% bedroeg. Daarnaast is voor de meeste kenmerken de genomic fokwaarde goed gecorreleerd met de pedigree-gebaseerde fokwaarde: voor de meeste kenmerken ligt de correlatie tussen de genomic fokwaarde en de pedigree-gebaseerde fokwaarde tussen 0.5 en 0.85, met uitzondering van het kenmerk levensduur (correlatie $r = 0.26$) en de kenmerken better life gezondheid en better life efficiëntie ($r < 0.05$). Op basis van de twee hierboven beschreven waarnemingen werd besloten om enkel de genomic fokwaarden te weerhouden als inputdata voor de machine-learning modellen. Aangezien de betrouwbaarheden voor een aantal belangrijke kenmerken (bv. vruchtbaarheid) ontbraken, werden deze niet weerhouden als inputdata voor de machine-learning modellen.

2.3.2 Geaggregeerde mpr-data

Het verwerken van de mpr-data tot een vast aantal variabelen met een duidelijk afgelijnde betekenis, vormde, onder andere door de grote variatie in het aantal beschikbare mpr's per lactatie, de grootste uitdaging van de primaire dataverwerking. Zoals eerder vermeld, waren voor iedere mpr 12 variabelen beschikbaar: pariteit, kalfdatum, proefmelkdatum, (on)geldigheid van de mpr, kg melk, % vet, % eiwit, %lactose, celgetal, ureumconcentratie, het aantal melkingen per dag en de status van de koe. Vooraleer deze mpr-data te verwerken, werden eerst alle afwijkende mpr-records verwijderd op basis van een aantal eenvoudige criteria, die in de volgende alinea besproken worden.

Betreffende de (on)geldigheid van de mpr's, waren er geen afgekeurde mpr's aanwezig, waardoor deze bijgevolg niet verwijderd moesten worden. Ook alle mpr's waarbij de koe een afwijkende status toegekend kreeg (tochtig, uierontsteking, monstername onmogelijk, vers gekalfd) werden opgezocht en uit de dataset verwijderd. Als derde weerhoudingscriterium werd vereist dat voor iedere mpr-record de productieresultaten (kg melk, % vet, % eiwit) groter waren dan 0. Daarnaast werden alle mpr-records waarbij een aantal melkingen per dag verschillend van 2, 3 of robot werd vermeld, ook niet weerhouden. Dit laatste weerhoudingscriterium houdt in dat ook mpr-records waarbij vermeld werd dat deze gebaseerd waren op eenmaal daags melken, werden verwijderd. Dit was initieel niet het plan, maar bij een verkennende analyse van de mpr-data, bleek het aantal mpr-records van koeien die maar eenmaal per dag werden gemolken te klein om deze te weerhouden in de datasets om de machine-learning modellen te trainen/evalueren. Als laatste weerhoudingscriterium op de individuele mpr-records, werd vereist dat de kalfdatum van de koe gekend moest zijn, zonder gekende kalfdatum is het immers niet mogelijk om op basis van de proefmelkdatum het aantal dagen in lactatie te berekenen. Tot slot werden nog twee weerhoudingscriteria

toegepast op dierniveau: dieren die tijdens hun leven op verschillende bedrijven melk geproduceerd hebben, werden verwijderd (gebaseerd op aan- en afvoergegevens) en er werd vereist dat de eerste mpr van iedere lactatie plaatsvond binnen de eerste 75 lactatiedagen.

Door de hierboven beschreven selectiecriteria toe te passen, werden enkel maar geldige mpr's weerhouden waarbij de koe een niet-afwijkende status toegekend kreeg. Hierdoor was geen betekenisvolle variatie meer aanwezig in de variabelen met betrekking tot de (on) geldigheid van de mpr en de status van de koe, waardoor deze verwijderd konden worden. Daarnaast konden de variabelen kalfdatum en proefmelkdatum samengevoegd worden tot een variabele lactatiedagen door het verschil van beiden te berekenen en te vermeerderen met één (een mpr meting die, hypothetisch, vlak na de kalving zou worden uitgevoerd, wordt immers tijdens dag 1 van de lactatie afgenomen en niet tijdens dag 0). De variabelen % lactose (slechts beperkte variabiliteit) en ureumconcentratie (relatief grote meetfout, waardoor deze enkel op groepsniveau geïnterpreteerd mogen worden) werden niet verder meegenomen bij het opstellen van de inputdatasets voor de machine-learning modellen. Verder werden de gehaltes (% vet en % eiwit) met behulp van de gemeten melkproductie (kg) omgerekend naar de dagproducties, uitgedrukt in kg vet en kg eiwit. Hierdoor werd het aantal variabelen herleid tot zeven: pariteit, lactatiedagen, kg melk, kg vet, kg eiwit, celgetal en aantal melkingen per dag.

Tot slot moest nog een manier gevonden worden om om te gaan met de variabiliteit in het aantal mpr's per lactatie en de variabiliteit in de frequentie waarmee mpr-data werden verzameld (variërend van iedere paar dagen tot iedere paar weken). Hierbij werd beoogd om een dataset (hierna geaggregeerde mpr-dataset genoemd) te verkrijgen met evenveel records als de originele dataset met mpr-gegevens, maar waarbij alle informatie van de vorige mpr's tijdens de lactatie is verwerkt in iedere record: de record van de eerste mpr bevat enkel informatie van de eerste mpr, de record van de tweede mpr bevat informatie van de eerste twee mpr's, ... en de record van de laatste mpr van een lactatie bevat informatie van alle mpr's van die lactatie. Om dit te bereiken werd voor elk van de zeven resterende variabelen een passende methode uitgewerkt, welke hierna wordt beschreven.

Pariteit, lactatiedagen en aantal mpr-metingen

De meest eenvoudige 'methode' is deze voor de pariteit, aangezien deze doorheen de volledige lactatie hetzelfde blijft, kan immers voor iedere record in de geaggregeerde mpr-dataset de pariteit in één enkele variabele worden weergegeven. De kolom met

betrekking tot de pariteit in de geaggregeerde mpr-dataset is in de geaggregeerde mpr-dataset bijgevolg identiek aan deze in de originele mpr-dataset.

Voor het aantal lactatiedagen werden vijf variabelen ingevoerd: het aantal lactatiedagen bij de laatst gekende mpr (dgn n), aangevuld met het aantal dagen in lactatie op het moment dat de eerste vier mpr-metingen werden uitgevoerd (dgn1, dgn 2, dgn 3, en dgn 4).

Deze zes variabelen werden nog aangevuld met een variabele die het aantal mpr-metingen bijhoudt dat reeds werd uitgevoerd tijdens de lactatie en vier dummievariabelen die aangeven welke van de eerste vier mpr-metingen reeds werden uitgevoerd (Tabel 2.2).

Tabel 2.2: Illustratie van het verloop van de variabelen met betrekking tot pariteit, aantal lactatiedagen en aantal mpr-metingen in de geaggregeerde mpr-dataset voor een hypothetische koe op een bedrijf met een mpr-frequentie van 4 weken

# mpr's	pariteit	mpr 1	dgn 1	mpr 2	dgn 2	mpr 3	dgn 3	mpr 4	dgn 4	mpr n	dgn n
1	1	1	10	0	0	0	0	0	0	1	10
2	1	1	10	1	38	0	0	0	0	2	38
3	1	1	10	1	38	1	66	0	0	3	66
4	1	1	10	1	38	1	66	1	94	4	94
5	1	1	10	1	38	1	66	1	94	5	122
6	1	1	10	1	38	1	66	1	94	6	150

Melkregime

Aangezien sommige bedrijven in de loop van de tijd veranderen van melkregime, is het aantal melkbeurten per dag waarop de mpr-metingen gebaseerd zijn (2, 3 of 9=robot) niet voor alle koeien een constante doorheen de tijd. Daarom werd er voor gekozen om zeven variabelen te gebruiken om het melkregime weer te geven in de geaggregeerde mpr-dataset. De eerste 4 variabelen geven de melkregimes van de eerste vier mpr's (indien beschikbaar) weer. De laatste 3 variabelen (*tweemaal daags*, *driemaal daags* en *robotmelken*) geven, vanaf het moment dat het aantal mpr-metingen groter is dan drie, de fractie van de mpr-metingen weer waarbij respectievelijk tweemaal daags, driemaal daags of met een robot werd gemolken. Een dier dat bij de eerste 5 mpr-metingen tweemaal daags werd gemolken, krijgt bij de record horende bij de vijfde mpr in de geaggregeerde mpr-dataset hierbij voor de variabele *tweemaal daags* dus 1 toegekend terwijl de variabelen *driemaal daags* en *robotmelken* beide de waarde 0 toegekend krijgen. Analoog krijgt een dier waarvoor

10 mpr-metingen beschikbaar zijn, waarvan 5 met tweemaal daags melken en 5 met robotmelken, respectievelijk 0.5, 0 en 0.5 toegekend voor de variabelen *tweemaal daags*, *driemaal daags* en *robotmelken*. Niet-beschikbare variabelen krijgen in de geaggregeerde mpr-dataset de waarde 0 toegekend. Naarmate het aantal beschikbare mpr-metingen gedurende de lactatie oploopt, krijgen dus steeds meer variabelen met betrekking tot melkregime een waarde verschillend van nul toegekend in de geaggregeerde mpr-dataset, wat wordt geïllustreerd in Tabel 2.3. Het primaire doel van de variabelen in Tabel 2.3 op te nemen in de dataset is om de machine-learning modellen toe te laten de melkproductiedata correct te interpreteren, eerder dan informatie te verschaffen over met welk melkregime het dier tijdens het vervolg van de lactatie zal gemolken worden.

Tabel 2.3: Illustratie van het verloop van de variabelen met betrekking tot het melkregime in de geaggregeerde mpr-dataset voor een koe die tijdens de eerste 2 mpr's driemaal daags werd gemolken en vanaf de derde mpr tweemaal daags werd gemolken

# mpr's	melkregime 1	melkregime 2	melkregime 3	melkregime 4	tweemaal daags	driemaal daags	robotmelken
1	3	0	0	0	0	0	0
2	3	3	0	0	0	0	0
3	3	3	2	0	0	0	0
4	3	3	2	2	0.5	0.5	0
5	3	3	2	2	0.4	0.6	0
6	3	3	2	2	0.33	0.67	0

Tot slot werden de melkregime-fracties horende bij de laatst gekende mpr-meting van ieder lactatie gekopieerd en bewaard in een aparte subdataset. Deze melkregime-fracties werden als variabelen beschouwd die reeds bij de geboorte van het dier gekend zijn.

Celgetal

Net zoals bij het melkregime, werden ook bij het celgetal de resultaten van de eerste vier mpr-metingen op zichzelf opgenomen in de geaggregeerde mpr-dataset (indien beschikbaar). Vanaf het moment dat het aantal mpr-metingen groter is dan drie, werden daarnaast nog zeven beschrijvende statistieken opgenomen in de geaggregeerde mpr-dataset: het gemiddelde, de standaardafwijking (std.), het minimum, het 25% percentiel, de mediaan, het 75% percentiel en het maximum. Er werd gewerkt met zeven beschrijvende statistieken om zo veel mogelijk informatie over de distri-

butie van de celgetallen te bewaren. Analoog aan het melkregime, zijn de beschrijvende statistieken voor iedere record in de geaggregeerde mpr-dataset enkel maar gebaseerd op de op dat moment reeds beschikbare mpr's: de beschrijving van de celgetal-distributie bij de record die overeenstemt met de zesde mpr van de lactatie, is dus gebaseerd op enkel en alleen de eerste zes celgetal-metingen en niet op de celgetalmetingen die eventueel nog volgen tijdens het verdere verloop van de lactatie.

Tabel 2.4: Illustratie van het verloop van de variabelen met betrekking tot het celgetal (×1000/ml) in de geaggregeerde mpr-dataset voor een hypothetische koe met, in chronologische volgorde, volgende celgetal-metingen (×1000/ml): 50, 100, 150, 200, 250, 300

# mpr's	celgetal 1	celgetal 2	celgetal 3	celgetal 4	gemiddelde	std.	minimum	25% percentiel	mediaan	75% percentiel	maximum
1	50	0	0	0	0	0	0	0	0	0	0
2	50	100	0	0	0	0	0	0	0	0	0
3	50	100	150	0	0	0	0	0	0	0	0
4	50	100	150	200	125	64.5	50	87.5	125	162.5	200
5	50	100	150	200	150	79.1	50	100	150	200	250
6	50	100	150	200	175	93.5	50	112.5	175	237.5	300

Kg melk, kg vet en kg eiwit

Nu blijven er nog 3 variabelen over: kg melk, kg vet en kg eiwit. Deze drie variabelen vertonen een relatief vast patroon (lactatiecurve) doorheen de tijd: na de kalving neemt het productieniveau snel toe tot een piekproductieniveau, waarna het productieniveau langzaam terug daalt. Aangezien het werken met zeven beschrijvende statistieken, zoals voor het celgetal werd gedaan, een groot deel van de informatie die vervat zit in het verloop van deze lactatiecurve onbeschikbaar zou maken voor de machine-learning modellen, werd een geavanceerdere aanpak gebruikt. Hierbij werd, indien mogelijk, een lactatiecurve-model gefit aan alle op dat moment beschikbare mpr-data en werden vervolgens de gefitte parameters opgenomen in de dataset. Aangezien de parameters het volledige functieverloop vastleggen, zit immers alle informatie met betrekking tot de lactatiecurve vervat in de parameterwaarden.

In de wetenschappelijke literatuur werden een aantal modellen opgezocht die lactatiecurves trachten te beschrijven (Tabel 2.5). Een eerste selectie werd gemaakt op basis het aantal modelparameters. Dit moest bij voorkeur zo klein mogelijk zijn aangezien, om unieke parameters te kunnen bepalen, er minstens evenveel mpr-metingen nodig zijn als er parameters zijn in het lactatiemodel. Een gevolg hiervan is dat alle

modellen die bij het literatuuronderzoek werden weerhouden (Tabel 2.5) een black-box structuur hebben. Mechanistische modellen, zoals die van Pollott (2000) hebben immers vaak een ingewikkeldere modelstructuur en meer parameters dan black-box modellen.

Tabel 2.5: Beschouwde lactatiemodellen

bron	**naam**	**# par.**	**model**
Wood (1967)	WD	3	$Y(t) = at^b e^{-ct}$
Yadav et al. (1977)	Y	3	$Y(t) = a + bt^{-1} + ct$
Guo en Swalve (1995)	GS3	3	$Y(t) = a + b\sqrt{t} + c\log t$
Guo en Swalve (1995)	GS4	4	$Y(t) = a + b\sqrt{t} + c\log t + dt^4$
Wilmink (1987)	WL	4	$Y(t) = a + be^{-kt} + ct$

De vijf modellen die werden weerhouden, werden geëvalueerd op basis van 2000 lactaties waarvoor de laatste mpr plaatsvond na 305 dagen in lactatie en op basis van 2000 (onvoltooide) lactaties waarbij de (voorlopig) laatste mpr voor dag 305 in lactatie plaatsvond. Voor de evaluatie zelf werden R^2-waardes berekend voor zowel de voorspelling van de dagproducties als de voorspelling van de 305d-productie. De predicties van de 305d-producties werden hierbij berekend door de gefitte lactatiemodellen te integreren van dag 0.5 tot dag 305.5. De referentiewaarnemingen waren hierbij de geregistreerde mpr-producties (kg melk, kg vet en kg eiwit) en de (voorspelde) 305-dagenproducties die werden aangeleverd door CRV (CRV, 1998). Voor het berekenen van de R^2-waardes werden alle predicties voor alle lactaties geaggregeerd in een vector $\hat{\boldsymbol{y}}$ en werden alle werkelijke waarden geaggregeerd in een vector $\boldsymbol{y}$, waarna op basis van deze twee vectoren R^2 werd berekend volgens formule (2.1).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{\boldsymbol{y}})^2} = 1 - \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2}{\left\|\boldsymbol{y} - \overline{\boldsymbol{y}}\right\|^2} \tag{2.1}$$

Tijdens de evaluatieprocedure werden de modellen enkel maar gefit op de onvolledige lactaties indien er minstens één mpr-waarneming meer aanwezig was dan het aantal parameters. Modellen WD, Y en GS3 werden dus gefit van zodra er 4 mpr-waarnemingen beschikbaar waren en modellen GS4 en WL van zodra er 5 waarnemingen beschikbaar waren. Indien dit kritisch aantal waarnemingen niet aanwezig was werden alle parameters, behalve de a-parameter gelijk gesteld aan 0. Voor alle modellen in Tabel 2.5 geeft dit aanleiding tot het model $Y(t) = \mu$, met μ het gemiddelde productieniveau van alle beschikbare metingen. Door dit zeer eenvoudige model te

gebruiken als het model in kwestie niet of enkel met volledige overfitting gefit kan worden, werd een complexiteits-penalty ingebouwd in de evaluatieprocedure. Voor model GS4 werd, door de grote gelijkenis met model GS3, licht van deze regel afgeweken: indien minder dan 3 waarnemingen aanwezig waren, werd het model $Y(t) = a$ gefit, indien 4 waarnemingen aanwezig waren werd model GS3 gefit en indien vijf of meer waarnemingen beschikbaar waren, werd model GS4 gefit. Praktisch wordt door het gebruik van deze complexiteits-penalty dus enkel model WL benadeeld bij de evaluatieprocedure.

Nog vooraleer de R^2-waarde berekend kon worden, kon reeds besloten worden dat model WL geen geschikt model was in het kader van deze masterproef. Dit omdat de standaard Python-methoden om curves te fitten zeer wispelturige parameterschattingen opleverden voor dit model en daarenboven de algoritmen er regelmatig zelfs niet in slaagden om parameters te bepalen binnen een aanvaardbaar tijdsbestek. Dit laatste gebeurde vooral bij lactaties die niet helemaal volgens een klassiek patroon verliepen. Model WL werd bijgevolg niet verder meegenomen in de evaluatieprocedure.

Model WD kon wel aan alle lactaties gefit worden, maar kan niet analytisch geïntegreerd worden. Dit is te omzeilen door numerieke integratie, maar geeft ook een indicatie dat het opnemen van enkel en alleen de parameters van model WD in de geaggregeerde mpr-dataset hoogstwaarschijnlijk niet tot hoogperformante machine-learning modellen ter predictie van productiekengetallen zal leiden, aangezien het verband tussen de parameters en eenvoudige productiekengetallen zoals de 305d-productie, uitgedrukt in kg melk, kg vet of kg eiwit, niet eens analytisch kan uitgedrukt worden.

De R^2-waarden die werden bekomen, worden weergegeven in Tabel 2.6. Het eerste wat in het oog springt in deze tabel zijn de negatieve R^2-waarden voor model GS4 met betrekking tot het voorspellen van 305d-producties indien de laatst gekende mpr werd afgenomen voor dag 305 in lactatie. Dit staat in scherp contrast met de hoge R^2-waarden die werden bekomen voor het modelleren van dagproducties voor diezelfde lactaties. De reden hiervoor is dat de laatste term van model GS4 (dt^4) er voor zorgt dat het model een heel stuk flexibeler is dan modellen GS3 en Y, maar dit leidt bij onvoltooide lactaties tot sterke overfitting en vaak een sterke, onrealistische, toe- of afname van de voorspelde melkproductie in het verder verloop van de lactatie, na de laatst gekende mpr. Deze verhoogde 'flexibiliteit' van model GS4 uit zich ook indien de R^2-waarden voor de dagproducties worden vergeleken tussen lactaties <305 dagen en lactaties ≥ 305 dagen. Voor modellen QD, Y en GS3 kan hierbij gezien worden dat de R^2-waarden voor onvolledige lactaties beduidend hoger zijn dan de overeenkomstige R^2-waarden voor afgeronde lactaties, wat te wijten is aan over-

fitting bij de onvolledige lactaties. De R^2-waarde voor model GS4 is daarentegen bij afgeronde lactaties nog steeds > 0.9, wat volledig te wijten is aan de extra 'flexibiliteit' van model GS4. Wordt gekeken naar het voorspellen van 305d-producties bij volledige lactaties, dan kan gezien worden dat de extra term in model GS4 weinig voordeel biedt: modellen GS3 en GS4 zijn nagenoeg equivalent als het op het voorspellen van 305d-producties bij volledige lactaties neerkomt. Model Y is wat betreft het voorspellen van 305d-producties steeds inferieur aan model GS3, ongeacht of het voltooide of onvoltooide lactaties betreft.

Tabel 2.6: Performantie (R^2) van verschillende lactatiemodellen bij het modelleren van dag- en 305d-producties

		dgn. < 305			**dgn. ≥ 305**		
	model	**kg M**	**kg V**	**kg E**	**kg M**	**kg V**	**kg E**
dag-prod.	WD	0.87	0.83	0.86	0.58	0.56	0.57
	Y	0.90	0.85	0.88	0.68	0.64	0.63
	GS3	0.90	0.85	0.88	0.68	0.62	0.63
	GS4	0.97	0.95	0.97	0.92	0.92	0.92
305d-prod.	WD	-	-	-	-	-	-
	Y	0.51	0.26	0.45	0.79	0.75	0.73
	GS3	0.80	0.75	0.80	0.98	0.97	0.97
	GS4	-6.58	-4.82	-2.41	0.97	0.96	0.97

Wat betreft de verschillende productiekenmerken (kg melk, kg vet en kg eiwit), kan in Tabel 2.6 gezien worden dat alle modellen voor de verschillende productiekenmerken steeds vrij gelijkaardige R^2-waardes realiseren. Bijgevolg is het dan ook niet nodig om voor de verschillende productiekenmerken verschillende modellen te kiezen.

Gegeven de hierboven beschreven waarnemingen dat (1) model GS3 beter 305d-producties kan voorspellen dan model Y en (2) dat model GS4 zich in het tijdsvenster na de laatste mpr-meting onrealistisch kan gedragen, lijkt model GS3 het meest geschikte model in het kader van deze masterproef.

Model GS3 heeft bovendien enkele bijkomende voordelen. Ten eerste is model GS3 lineair in zijn parameters. Hierdoor kan lineaire regressie gebruikt worden om de modelparameters te bepalen, waardoor de benodigde rekentijd beperkt blijft. Een tweede voordeel van deze lineariteit in de parameters is dat productieparameters die te maken hebben met het productieniveau op een bepaald tijdstip (bv. de melkproductie op dag 30/60/90/...), te maken hebben met integratie van de lactatiecurve over een vast tijdsinterval (bv. productie tijdens de eerste 60/100/200/305 dagen) of te maken hebben met afgeleiden van de lactatiecurve (bv. snelheid waarmee de melkproductie toe- of afneemt rond dag 10/20/30...) niet moeten worden opgenomen in de inputdataset, aangezien deze stuk voor stuk worden gegeven door een eenvoudige lineaire combinatie van de parameters van model GS3.

Net zoals bij het celgetal en het aantal melkingen per dag, worden ook voor de productie van melk, vet en eiwit, de resultaten van de eerste vier mpr-metingen als individuele variabelen opgenomen in de geaggregeerde mpr-dataset (zie Tabel 2.7). Vanaf het moment dat 4 mpr-metingen beschikbaar zijn, worden daarnaast ook de parameters van het gefitte model GS3 opgenomen in de dataset.

Tabel 2.7: Illustratie van het verloop van de variabelen met betrekking tot de productie, uitgedrukt in kg melk (kg M), kg vet (kg V) en kg eiwit (kg E) in de geaggregeerde mpr-dataset. De data die worden weergegeven zijn afkomstig van een willekeurig geselecteerde koe

	# mpr's	mpr 1	mpr 2	mpr 3	mpr 4	a	b	c
kg M	**1**	28.9	0	0	0	0	0	0
	2	28.9	35.1	0	0	0	0	0
	3	28.9	35.1	35.4	0	0	0	0
	4	28.9	35.1	35.4	25.7	3.4	-9.1	24.6
	5	28.9	35.1	35.4	25.7	10.9	-6.0	16.9
	6	28.9	35.1	35.4	25.7	14.9	-4.6	13.2
kg V	**1**	1.21	0	0	0	0	0	0
	2	1.21	1.45	0	0	0	0	0
	3	1.21	1.45	1.44	0	0	0	0
	4	1.21	1.45	1.44	1.02	0.15	-0.40	1.04
	5	1.21	1.45	1.44	1.02	0.52	-0.24	0.67
	6	1.21	1.45	1.44	1.02	0.70	-0.18	0.49
kg E	**1**	1.08	0	0	0	0	0	0
	2	1.08	1.25	0	0	0	0	0
	3	1.08	1.25	1.23	0	0	0	0
	4	1.08	1.25	1.23	0.88	0.24	-0.32	0.84
	5	1.08	1.25	1.23	0.88	0.53	-0.20	0.54
	6	1.08	1.25	1.23	0.88	0.66	-0.16	0.42

2.3.3 305d- en lactatieproducties

In §2.3.2 werd aangegeven dat alle productieparameters die verband houden met het productieniveau op een vast tijdstip, integratie van de lactatiecurve over een vast tijdsinterval en afgeleiden van de lactatiecurve, niet moeten worden opgenomen in de datasets om de machine-learning modellen te trainen en te evalueren. De totale lactatieproductie vormt echter een belangrijke uitzondering op deze regel, aangezien hierbij geïntegreerd moet worden over een variabel tijdsinterval (tussen 0.5 en de lactatielengte + 0.5). Om deze reden werden de totale lactatieproducties, uitgedrukt in kg melk, kg vet en kg ewit, uitgerekend op basis van de lactatiecurve die gefit werd aan alle mpr-metingen van een bepaalde lactatie (de parameters van deze lactatiecurve worden in de geaggregeerde mpr-dataset opgenomen bij de laatst gekende mpr-meting van iedere lactatie). Hierdoor konden de lactatieproducties, indien relevant, ook toegevoegd worden aan de datasets om de machine-learning modellen te

trainen/evalueren. Aangezien de 305d-productie-voorspellingen van model GS3 niet volledig overeenkwamen met deze die door CRV werden aangeleverd, werden ook de 305d-producties uit de ruwe data toegevoegd aan de subdataset met lactatieproducties. Dit opnieuw met als doel om deze productieparameters, indien relevant, te kunnen toevoegen aan de datasets om de machine-learning modellen te trainen/evalueren.

2.3.4 Vruchtbaarheidsdata

Analoog aan de mpr-data is er een grote verscheidenheid in het aantal conceptiepogingen dat nodig is om een koe drachtig te krijgen en het tijdstip waarop deze plaatsvinden. Daarom werden de reproductie-activiteiten, vrij analoog aan hoe dit voor de mpr-data gebeurde, omgezet in een vast aantal variabelen met een eenduidige betekenis. Vooraleer dit te doen, werden eerst alle dieren verwijderd die ooit werden gespoeld, samengeweid met een stier of waarbij een embryo werd ingeplant. Vervolgens werd alle informatie van de conceptiepogingen (kunstmatige inseminatie of natuurlijke dekking) binnen eenzelfde lactatie geaggregeerd in 4 variabelen: conceptiepoging 1 beschikbaar (*ins 1*), leeftijd/lactatiedagen bij eerste conceptiepoging (*dgn 1*), totaal aantal conceptiepogingen (*ins n*) en leeftijd/lactatiedagen bij de laatst gekende conceptiepoging (*dgn n*). Hierbij werd per conceptiepoging een record gecreëerd, waarbij voor de berekening van de hiervoor beschreven variabelen, alle op dat moment reeds beschikbare informatie omtrent conceptiepogingen werd gebruikt. Per pariteit bevat de eerste record dus enkel informatie over de eerste conceptiepoging, de tweede record informatie over de eerste twee conceptiepogingen,... en de laatste record informatie over alle conceptiepogingen tijdens die pariteit. Een voorbeeld van hoe deze geaggregeerde data er uit kunnen zien voor de pariteiten nul tot en met twee, is terug te vinden in Tabel 2.8. Naast informatie over conceptiepogingen, leverde CRV ook data aan met betrekking tot drachtigheidscontroles. Aangezien een niet verwaarloosbaar deel van de dieren na een positieve drachtcontrole toch opnieuw een of meerdere conceptiepoging(en) onderging, werden deze data niet verder meegenomen bij het opstellen van de datasets om de machine-learning modellen te trainen en evalueren.

2.3.5 Epigenetische effecten

Slechts twee epigenetica-gerelateerde factoren konden uit de ruwe data worden geïsoleerd: de pariteit van de moeder en het geboortegewicht. Met betreking tot de pariteit van de moeder, werden vijf categoriën onderscheiden: 0, 1, 2, ≥3 en on-

Tabel 2.8: Illustratie van het verloop van de variabelen met betrekking tot vruchtbaarheid in de geaggregeerde mpr-dataset. In de tabel worden gegevens weergegeven van een willekeurig geselecteerde koe

pariteit	datum	ins 1	dgn 1	ins n	dgn n
0	14/10/2006	1	403	1	403
1	14/10/2007	1	89	1	89
2	01/11/2008	1	101	1	101
2	20/11/2008	1	101	2	120
2	06/01/2009	1	101	3	167

bekend. Er werd geopteerd om met een vijfde catagorie te werken indien de pariteit van de moeder ongekend was, aangezien het niet weerhouden van alle dieren waarvoor de pariteit van de moeder ongekend was, tot een te grote reductie van het aantal beschikbare records zou geleid hebben. Naast de pariteit van de moeder werd ook het geboortegewicht meegenomen in de inputdatasets voor de modellen. Omdat ook voor het geboortegewicht het niet weerhouden van alle dieren zonder gekend geboortegewicht zou geleid hebben tot een te grote reductie van het aantal beschikbare records, kregen dieren waarvoor het geboortegewicht onbekend was een default-geboortegewicht toegekend. De gebruikte defaultwaarde werd hierbij berekend als het gemiddelde van alle gekende geboortegewichten van vaarskalveren. Daarnaast kregen ook alle dieren met een geboortegewicht lager dan 10 kg het default-geboortegewicht toegekend, aangezien deze geboortegewichten met nagenoeg 100% zekerheid aan fouten bij de registratie van het geboortegewicht kunnen gewijd worden. Met betrekking tot meerlingen, leverde CRV (indien gekend) de verschillende geboortegewichten van de verschillende kalveren aan, maar was niet opgegeven welk geboortegewicht bij welk kalf hoorde. Daarom werd voor meerlingen aan alle kalveren van de meerling het gemiddelde van de geregistreerde geboortegewichten toegekend.

2.3.6 Levensverloop

Naast gegevens over de productie en reproductie van de individuele dieren, werd uit de verschillende deeldatasets die werden aangeleverd door CRV, ook het levensverloop geëxtraheerd. Het levensverloop werd hierbij beschouwd als een reeks datums die aangeven wanneer een bepaalde pariteit start, wanneer de koe wordt drooggezet en wanneer de pariteit eindigt. Pariteiten werden hierbij als beëindigd beschouwd indien het dier (opnieuw) afkalfde of werd afgevoerd.

2.3.7 Bedrijfsparameters

Naast de primaire dataverwerking van de ruwe data met betrekking tot productie en reproductie op dierniveau (§2.3.4 en §2.3.2) werden tijdens de primaire data-analyse ook bedrijfsparameters berekend met betrekking tot productie, reproductie en vervangingsbeleid. Analoog aan wat gedaan werd bij de primaire dataverwerking van de mpr-gegevens, werden ook hier eerst alle dieren verwijderd waarbij niet alle mpr-metingen op hetzelfde bedrijf werden uitgevoerd.

Productieparameters

De productie-gerelateerde bedrijfsparameters werden voor iedere eerste dag van de maand berekend. Naast het celgetal en het melkregime, werden vijf 305d-productiekengetallen beschouwd: kg melk (kg M), kg vet (kg V), kg eiwit (kg E), percentage vet (%V) en percentage eiwit (%E). Om jaarinvloeden te vermijden, werd voor de berekeningen gewerkt met een tijdsvenster van twee jaar: voor het celgetal en het melkregime werd gebruik gemaakt van mpr-data die werden verzameld in de twee jaar voorafgaand aan de berekeningsdatum. Voor kg M, kg V, kg E, %V en %E werden alle lactaties beschouwd die werden afgerond in de twee jaar voor de berekeningsdatum. Voor het melkregime werden 3 bedrijfsparameters berekend, die elk de fractie van de mpr-metingen weergeven waarbij respectievelijk tweemaal daags, driemaal daags of met een robot werd gemolken. Voor alle andere productie-gerelateerde parameters werd het gemiddelde, de standaardafwijking, het 25 % percentiel, de mediaan en het 75 % percentiel berekend, dit om zo veel mogelijk informatie met betrekking tot de distributie van de productieparameters binnen het bedrijf te bewaren. Hierbij werden de koeien ingedeeld in drie pariteitsgroepen (1, 2 en $\geq$3), waarbij de hiervoor genoemde statistieken zowel op bedrijfsniveau als voor iedere pariteitsgroep apart werden berekend. Indien een van deze drie pariteitsgroepen voor een bepaalde berekeningsdatum minder dan 10 dieren bevatte, werd deze berekeningsdatum niet weerhouden.

Reproductieparameters

De reproductie-gerelateerde bedrijfsparameters werden op een analoge manier benaderd als de productie-gerelateerde bedrijfsparameters: de bedrijfsparameters werden voor iedere eerste dag van de maand berekend en alle pariteiten die werden afgerond binnen een tijdsvenster van twee jaar voor de berekeningsdatum werden meegenomen in de berekeningen. Bovendien werd ook gewerkt met drie pariteitsgroepen (0 = pink, 1 = vaars en $\geq$2 = koe). Analoog zoals gebeurde op individueel

dierniveau, werden dieren die in hun leven gespoeld werden, een embryo implantatie ondergingen of werden samengeweid met een stier, niet weerhouden voor het berekenen van de bedrijfsparameters met betrekking tot reproductie. Verder werden enkel pariteiten die werden afgesloten met een kalving weerhouden, dit omdat er op nagenoeg ieder bedrijf wel een berekeningsmoment was waarop een koe die veelvuldig geïnsemineerd werd, maar toch niet drachtig werd, de bedrijfsparameters sterk beïnvloedde. Daarnaast werden bij de pariteiten 1 en ≥2 ook alle pariteiten verwijderd waarbij de eerste inseminatie niet tussen dag 21 en 365 van de lactatie plaatsvond of waarbij de tussenkalftijd korter was dan 280 dagen (bv. ten gevolge van een abortus). De reproductiekengetallen die werden beschouwd zijn de leeftijd bij eerste dekking (par. 0) het interval afkalven-eerste inseminatie (par. 1 en ≥2) het aantal dekkingen per dracht, de leeftijd bij eerste kalving (par. 0) en de tussenkalftijd (par. 1 en ≥2). Analoog aan de productie-gerelateerde bedrijfsparameters, werden ook voor de reproductie-gerelateerde bedrijfsparameters per pariteitsgroep verschillende statistieken berekend om maximale informatie te behouden over de distributie van de reproductiekengetallen binnen het bedrijf: het gemiddelde, de standaardafwijking en het 25%, 50% en 75% percentiel. Indien voor een van de drie beschouwde pariteitsgroepen voor minder dan 10 dieren gegevens beschikbaar waren, werd de berekeningsdatum voor het bedrijf in kwestie niet weerhouden. Daarnaast werden ook nog vier bedrijfsparameters opgesteld met betrekking tot de toepassing van natuurlijke dekking: gebruik bij pinken (0/1), aandeel op het totale aantal conceptiepogingen bij pinken, gebruik bij vaarzen en koeien (0/1) en aandeel op het totale aantal conceptiepogingen bij vaarzen en koeien.

Vervangingsbeleid

Ook voor parameters met betrekking tot het vervangingsbeleid, zoals het vervangingspercentage, was het initiële idee om deze parameters voor iedere eerste dag van de maand te berekenen. Specifiek met betrekking tot het vervangingspercentage zou hierbij formule (2.2) gebruikt worden, met RR het vervangingspercentage, n_1 het aantal koeien aanwezig een jaar voor de berekeningsdatum, n_2 het aantal koeien aanwezig op de berekeningsdatum en c het aantal dieren dat werd afgevoerd in de loop van het jaar voor de berekeningsdatum.

$$RR = \frac{2\,c}{n_1 + n_2} \tag{2.2}$$

Echter, een maandelijkse berekening van het vervangingspercentage met formule (2.2) leidde er toe dat een onrealistisch groot deel van de bedrijven vervangingspercentages noteerde lager dan 20% voor berekeningsmomenten vroeger dan 2013-

2014 De reden hiervoor is waarschijnlijk de manier waarop CRV de dieren selecteerde die werden opgenomen in de dataset, waardoor een deel van de dieren die voor 2013-2014 werd afgevoerd, niet werd opgenomen in de dataset. Voor de productie- en reproductie-gerelateerde kenmerken vormde dit geen probleem (geen opvallende verschillen in de distributie van de bedrijfsparameters voor en na 2013-2014), maar bij het maandelijks (her)berekenen van het vervangingspercentage, leidde dit tot een duidelijke neerwaartse bias voor veel bedrijven in de periode voor 2013-2014. Aangezien het vervangingspercentage een belangrijke bedrijfsparameter is die niet in de inputdatasets voor de machine-learning modellen mocht ontbreken en deze bedrijfsparameter ook voor de jaartallen voor 2014 gekend moest zijn, werd per bedrijf een enkel meerjarig vervangingspercentage berekend. Hiervoor werd gebruik gemaakt van gegevens omtrent het aantal aanwezige en afgevoerde dieren in de periode van 1 januari 2014 tot 1 januari 2019. Om dit meerjarig vervangingspercentage te berekenen werd formule (2.2) licht aangepast tot formule (2.3), met n_i het aantal dieren aanwezig op 1 januari van jaar i en c_i het aantal afgevoerde dieren in de loop van jaar i. Deze lichte aanpassing werd doorgevoerd om te verzekeren dat ieder afgevoerd dier even zwaar zou meetellen bij de berekening van het meerjarig vervangingspercentage. Bedrijven die tussen 2014 en 2018 minder dan 50 dieren afvoerden, werden niet weerhouden bij het berekenen van de vervangingspercentages.

$$RR = \frac{2 \sum_{i=2014}^{2018} (c_i)}{\sum_{i=2014}^{2018} (n_i + n_{i+1})} \tag{2.3}$$

Naast het vervangingspercentage werd ook de pariteitsdistributie berekend van de dieren die werden afgevoerd in de periode 2014-2018. Hierbij werd gewerkt met 6 pariteitscategorieën (1, 2, 3, 4, 5 en ≥6) en werd voor iedere pariteitscategorie een bedrijfsparameter berekend die het aandeel van een bepaalde pariteitscategorie in het totaal aantal afgevoerde dieren uitdrukt.

2.3.8 Productiekengetallen

Naast het berekenen van allerlei inputvariabelen voor de machine-learning modellen, werden ook de te voorspellen productiekengetallen geïsoleerd uit de ruwe data (Tabel 2.9). Hierbij werden drie categroriën van productiekengetallen beschouwd: gelinkt aan de eerste lactatie, gelinkt aan de productie op x jaar leeftijd en gelinkt aan de levensproductie.

Met betrekking tot de productiekengetallen gelinkt aan de eerste lactatie, werd gefocust op kengetallen die eenvoudig bepaald konden worden op basis van de 305d-producties die werden aangeleverd door CRV: kg melk (kg M), kg vet (kg V), kg eiwit

(kg E), kg vet+eiwit (kg V+E), percentage vet (%V), percentage eiwit (%E), kg Fat Corrected Milk (kg FCM) en kg Fat and Protein Corrected Milk (kg FPCM). Voor de berekening van de 305d-productie, uitgedrukt in kg FCM en kg FPCM, werd gebruik gemaakt van formules (2.4) en (2.5) (CVB, 2016).

$$\begin{aligned} \text{FCM} &= (0.4 + 0.15 \times \%\text{V}) \times \text{kgM} \\ &= 0.4 \times \text{kgM} + 15 \times \text{kgV} \end{aligned} \tag{2.4}$$

$$\begin{aligned} \text{FPCM} &= (0.337 + 0.116 \times \%V + 0.06 \times \%\text{E}) \times \text{kgM} \\ &= 0.337 \times \text{kgM} + 11.6 \times \text{kgV} + 6 \times \text{kgE} \end{aligned} \tag{2.5}$$

Aangezien het weinig extra werk vroeg om naast de 305d-productie ook de tussenkalftijd (tkt) tussen de eerste en de tweede pariteit te berekenen op basis van de data die werden aangeleverd, werd ook de tkt meegenomen als kengetal gelinkt aan de eerste lactatie. De tkt is weliswaar een vruchtbaarheidskengetal en geen productiekengetal, maar aangezien het meenemen van de tkt als extra kengetal informatie kan verschaffen over het potentieel van machine-learning technieken bij het voorspellen van vruchtbaarheidskengetallen, leek het wel interessant om ook dit kengetal mee te nemen bij het opstellen en analyseren van de machine-learning modellen.

Tabel 2.9: Overzicht van de kengetallen die werden berekend op basis van de aangeleverde data. In de tabel wordt voor ieder kengetal het gemiddelde (μ) en de standaardafwijking (σ) weergegeven

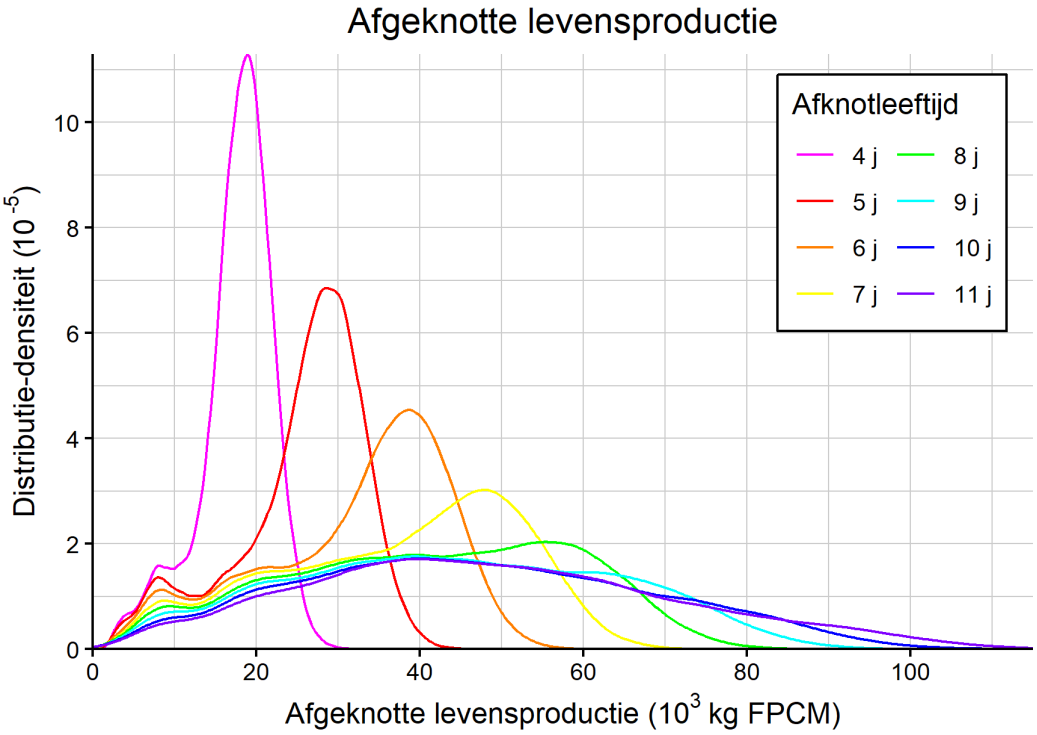
groep	**kengetal**	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$
1e lactatie	305d kg M	8259	1383
	305d kg V	349	52
	305d kg E	289	45
	305d kg V+E	638	93
	305d %V	4.26	0.46
	305d %E	3.51	0.21
	305d kg FCM	8540	1261
	305d kg FPCM	8567	1265
	tkt	398	70
x jaar lft.	kg FPCM; 3 jaar	8788	1921
	kg FPCM; 4 jaar	18857	3156
	kg FPCM; 5 jaar	29296	4282
	kg FPCM; 6 jaar	39745	5361
	kg FPCM; 7 jaar	50181	6427
	kg FPCM; 8 jaar	60464	7484
leven	kg FPCM	44482	19429

Voor de productiekengetallen gelinkt aan de productie op x jaar leeftijd, werd gekozen om de FPCM-productie te beschouwen die het dier realiseerde op x jaar leeftijd, met

$x \in \{3, 4, 5, 6, 7, 8\}$ De keuze voor FPCM werd hierbij gemaakt op basis van het feit dat FPCM een zeer algemeen kengetal is: zowel de hoeveelheid geproduceerde melk (∼ lactose) als de hoeveelheid geproduceerd vet en eiwit worden meegenomen in de berekening. Aangezien de FPCM-productie op x jaar leeftijd niet werd aangeleverd door CRV, werd dit productiekengetal zelf berekend op basis van de eerder opgestelde geaggregeerde mpr-dataset (§2.3.2). Hierbij werd eerst de datum bepaald waarop een dier x jaar werd, waarna voor alle begonnen lactaties de lactatiecurve-parameters werden opgehaald uit de geaggregeerde mpr-dataset. Er werd steeds gebruik gemaakt van de lactatiecurve-parameters horende bij de laatste mpr van de lactatie, aangezien deze gebaseerd zijn op alle mpr-metingen die tijdens de betreffende lactatie werden uitgevoerd. Vervolgens werden de lactatiecurves op een passende manier geïntegreerd, waarna na sommatie van deze integralen per productiekenmerk (kg M, kg V en kg E) en gebruik van formule (2.5), de FPCM-productie op x jaar leeftijd kon worden bepaald.

Ook voor de levensproductie werd er voor gekozen om deze uit te drukken in kg FPCM. Analoog aan de productie op x jaar leeftijd, werd de levensproductie hierbij berekend door de lactatiecurves van alle begonnen lactaties te integreren over hun volledige lengte, vervolgens deze integralen te sommeren per kenmerk (kg M, kg V en kg E) en tot slot formule (2.5) toe te passen. Echter, indien de levensproductie zou berekend worden voor alle dieren in de dataset waarvoor de afvoerdatum gekend is, zou dit er toe leiden dat langlevende dieren ondervertegenwoordigd zouden zijn in de waarnemingen, aangezien een deel van deze dieren nog in leven is. Dit zou ongetwijfeld tot een neerwaartse bias op de voorspellingen van de levensproducties geleid hebben, wat vermeden moest worden. Langs de andere kant was enkel en alleen data gebruiken van dieren die geboren werden in een geboortejaar waarvan reeds alle dieren waren afgevoerd ook geen optie, aangezien er dan amper bruikbare records zouden overblijven. Daarom werd gebruik gemaakt van een compromis-oplossing, waarbij de levensproducties werden 'afgeknot' op x jaar leeftijd. Deze afgeknotte levensproductie is gelijk aan de werkelijke levensproductie, tenzij het dier ouder werd dan x jaar, dan wordt de afgeknotte levensproductie gelijk gesteld aan de productie op x jaar leeftijd. Door bovendien enkel maar data te gebruiken van dieren die op 31 december 2019 de leeftijd van x jaar konden bereiken, kon er een evenwichtige reeks aan (afgeknotte) levensproducties worden opgesteld, waarbij langlevende dieren niet ondervertegenwoordigd waren in de metingen.

Om de juiste waarde van x te bepalen, werd voor verschillende waarden van x de afgeknotte levensproductie bepaald van alle dieren die op 31 december 2019 de leeftijd van x jaar konden bereiken, waarna de distributie van deze afgeknotte levensproducties werd bestudeerd. Op Figuur 2.1 kan gezien worden dat voor $x \leq 8$ jaar



Figuur 2.1: Distributie van de afgeknotte levensproductie voor verschillende mogelijke waarden van de afknotleeftijd

een duidelijk storende invloed van de afknotting aanwezig is het rechterdeel van de distributie. Vanaf $x = 9$ jaar, is echter geen storende invloed van de afknotting meer zichtbaar in de distributie. De levensproductie afknotten op 10 of 11 jaar leeftijd is weliswaar nog beter, maar aangezien dit ten koste gaat van het aantal dieren dat kan worden meegenomen in de datasets om de machine-learning modellen te trainen en te evalueren, werd gekozen om een afknotleeftijd van 9 jaar te hanteren bij het berekenen van de (afgeknotte) levensproducties. In alles wat volgt zal terug de term 'levensproductie' gebruikt worden, maar wees indachtig dat dit hiermee steeds de 'afgeknotte levensproductie' bedoeld zal worden.

2.3.9 Interactie tussen fokwaarden en bedrijf

In een klassiek lineair model $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \boldsymbol{\epsilon}$ wordt een interactie-effect tussen een variabele $\boldsymbol{x_1}$ en een variabele $\boldsymbol{x_2}$ doorgaans in het model in rekening gebracht door een term in functie van $\boldsymbol{x_1 x_2}$ op te nemen in het model: $\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \beta_3 \boldsymbol{x_1 x_2} + \boldsymbol{\epsilon}$. Voor het inbrengen van het interactie-effect tussen de fokwaarden en de bedrijfsomstandigheden in de machine-learning modellen, bleek deze klassieke aanpak echter om verschillende reden niet geschikt te zijn. Een eerste reden is het aantal extra variabelen dat zou moeten ingebracht worden in het mo-

del: met 74 gekende fokwaarden en ± 200 verschillende bedrijven, zou het inbrengen van een interactie-effect tussen de fokwaarden en de bedrijven al snel leiden tot $(74-1)*(200-1)=14527$ extra variabelen in de inputdatasets. De meeste machine-learning technieken kunnen hier wel mee om, maar dit zou onnodig veel extra complexiteit met zich meebrengen. Een tweede reden is dat we graag de mogelijkheid wilden behouden om de machine-learning modellen te evalueren op basis van data van bedrijven die niet werden opgenomen in de trainingsdataset. Deze wens sluit het gebruik van de klassieke methodologie om rekening te houden met interactie-effecten helemaal uit, aangezien deze niet in staat is om een fokwaarde × bedrijf interactie-effect te bepalen voor bedrijven die niet in de trainingsdataset zijn opgenomen. Tot slot wilden we graag over de mogelijkheid beschikken om bij het opstellen van machine-learning modellen met betrekking tot productiekengetal a ook rekening te kunnen houden met het fokwaarde × bedrijf interactie-effect van een ander productiekengetal b, aangezien dit laatste interactie-effect andere, maar mogelijks relevante, informatie bevat over het bedrijfsmanagement. Ook hiervoor is de klassieke methodologie niet geschikt. Omwille van de hiervoor beschreven drie redenen werd een alternatieve 2-staps methodologie uitgewerkt om het interactie-effect tussen de fokwaarden en de bedrijven mee te kunnen nemen in de modellen.

Tijdens de eerste stap van deze alternatieve methodologie om rekening te houden met het fokwaarde × bedrijf interactie-effect, werd gestart met alle fokwaarden en productiekengetallen te standaardiseren ($\mu=0$, $\sigma^2=1$) op landelijk niveau. Daarna werd voor iedere unieke combinatie van een productiekengetal $i \in \{1,\ldots,n_i\}$, een fokwaarde $j \in \{1,\ldots,n_j\}$ en een bedrijfsnummer $k \in \{1,\ldots,n_k\}$ een regressie-analyse van productiekengetal i op fokwaarde j uitgevoerd (vgl. 2.6). Hierbij werden de productiekengetallen en fokwaarden van alle dieren (index l) die op bedrijf k hun volledige productieve leven hebben doorgebracht, gebruikt om de trainingsdatasets voor deze enkelvoudige lineaire regressiemodellen op te stellen.

$$y_{ijkl} = a_{ijk} + b_{ijk}x_{ijkl} + \epsilon_{ijkl} \tag{2.6}$$

Het achterliggende idee hiervan is dat de regressiecoëfficiënten b_{ijk} vervolgens per productiekengetal i kunnen ondergebracht worden in n_j variabelen (één per fokwaarde) die gebruikt kunnen worden om het interactie-effect tussen de fokwaarden en de bedrijven in rekening te brengen in de modellen, zonder dat hierbij de beperking ontstaat dat het opgestelde model enkel kan gebruikt worden om voorspellingen te maken voor bedrijven die werden opgenomen in de trainingsdataset. Voor bedrijven die niet opgenomen werden in de trainingsdataset, volstaat het dan immers om de betreffende regressiecoëfficiënten b_{ijk} te bepalen en deze, na toepassen van de hierna beschreven tweede stap, te gebruiken als inputs voor de machine-learning modellen.

Bovendien hebben de bekomen regressiecoëfficiënten ook enige praktijkwaarde indien het fokdoel perfect samenvalt met een van de 16 beschouwde kengetallen (bv. een zo hoog mogelijke levensproductie). In dat geval geven zij immers de gewichten aan die de verschillende fokwaarden in het paringsprogramma moeten krijgen om zo snel mogelijk een zo groot mogelijke vooruitgang te boeken in het fokdoel.

Omdat het voor verschillende productiekengetallen (bv. levensproductie) weinig zinvol is om de regressiecoëfficiënten b_{ijk} maandelijks te (her)berekenen zoals voor de productie- en reproductie gerelateerde bedrijfsparameters in §2.3.7 werd gedaan, werd voor de hele periode die de dataset bestrijkt per unieke combinatie van i, j en k slechts eenmaal model (2.6) gefit. Hierbij werd als weerhoudingscriterium gesteld dat van minimaal 10 dieren de benodigde informatie (fokwaarden + productiekengetal i) beschikbaar moest zijn.

Concreet leverde deze eerste stap in het kader van deze masterproef dus voor de 16 beschouwde productiekengetallen telkens 74 variabelen op die konden gebruikt worden om het fokwaarde × bedrijf interactie-effect mee te nemen in de machine-learning modellen. Indien het fokwaarde × bedrijf interactie-effect van alle productiekengetallen tegelijk in de inputdataset zou worden opgenomen (wat gedaan werd bij het modelleren van de levensproductie) hield dit echter nog steeds in dat er 1184 extra variabelen moesten worden opgenomen in de datasets om de machine-learning modellen te trainen en te evalueren. Daarom werd tijdens een tweede stap met behulp van principale componenten analyse het aantal variabelen per productiekengetal verder herleid van 74 naar 10. Hiervoor werd per productiekengetal principale componenten analyse toegepast op de 74 regressiecoëfficiënt-variabelen (een per fokwaarde) die bij dit productiekengetal hoorden, waarna enkel de eerste 10 principale componenten werden weerhouden. Het aandeel van de totale variatie in de 74 regressiecoëfficiënt-variabelen dat werd verklaard door deze eerste 10 principale componenten, varieerde afhankelijk van het kenmerk tussen 62% (kg FPCM op 8 jaar lft.) en 79% (305d kg E 1^{e} lactatie).

HOOFDSTUK 3

VOORSPELLEN VAN PRODUCTIEKENGETALLEN

De literatuurstudie en de primaire dataverwerking hadden tot doel respectievelijk kennis en consistente data aan te leveren, zodat op een doordachte manier degelijke machine learning modellen konden worden opgesteld om productiekengetallen van melkvee te voorspellen. In dit hoofdstuk wordt eerst verduidelijkt hoe de effectieve inputdatasets voor de machine learning modellen werden opgesteld, vertrekkende van de verschillende subdatasets die werden gecreëerd tijdens de primaire dataverwerking. Vervolgens wordt besproken welke machine learning technieken werden beschouwd en hoe machine learning modellen werden opgesteld (getraind) en geëvalueerd. Daarna worden de evaluatieresultaten van de opgestelde machine learning modellen uitvoerig besproken, om te eindigen met het bespreken van een mogelijke praktijktoepassing: het afvoeren van overtollig jongvee.

3.1 Opstellen van de inputdatasets

In hoofdstuk 2 werd besproken hoe de ruwe data werden verwerkt tot een aantal subdatasets (fokwaarden, geaggregeerde mpr-data, vruchtbaarheidsdata,...) die elk een deel van de informatie aanwezig in de ruwe data trachten te omvatten in variabelen met een duidelijk afgelijnde betekenis. Hierbij werd veel belang gehecht aan het aggregeren van de data, zodat, ongeacht het aantal uitgevoerde mpr-metingen of conceptiepogingen, alle beschikbare informatie steeds kon worden weergegeven in een vast aantal variabelen.

Vooraleer de machine learning modellen effectief getraind konden worden, moest de informatie aanwezig in deze subdatasets echter eerst nog op de juiste manier geassembleerd worden tot de effectieve inputdatasets voor de machine learning modellen. De eerste stap hierbij was het berekenen van de referentiedatum voor ieder dier: de datum waarop een dier het beschouwde referentiemoment (bv. twee maanden na de

geboorte) bereikte. Dieren waarvoor de referentiedatum niet berekend kon worden, werden niet weerhouden in de inputdataset.

Eenmaal de referentiedatum gekend was, was het vrij eenvoudig om alle data te assembleren. Eerst werd alle referentiedatum-onafhankelijke data aan de dataset toegevoegd: fokwaarden, fokwaarde × bedrijf interactie-effect, epigenetische factoren en melkregime van alle lactaties die gestart werden vooraleer het productiekengetal werd gerealiseerd. De fokwaarde × bedrijf interactie-effecten werden hierbij hiërarchisch toegevoegd (*1^e lactatie < x jaar lft. < leven*). Dit houdt in dat voor een gegeven productiekengetal, alle interactie-effecten met betrekking tot productiekengetallengroepen die zich op een lager of gelijk hiërarchisch niveau bevinden, toegevoegd worden aan de inputdataset. Voor het modelleren van de levensproductie werden dus alle interactie-effecten opgenomen in de inputdataset, terwijl voor het modelleren van kengetallen van groep *1^e lactatie* enkel de interactie-effecten van de kengetallen van groep *1^e lactatie* werden opgenomen. De informatie over melkregimes van alle gestarte lactaties werd door uitmiddeling herleid tot drie fracties (tweemaal daags melken, driemaal daags melken en robotmelken) om alle informatie over het levensverloop uit deze variabelen te elimineren. Dieren waarvoor een deel van de referentiedatum-onafhankelijke informatie niet beschikbaar was, werden niet weerhouden in de dataset.

Nadat alle referentiedatum-onafhankelijke informatie was toegevoegd, werd alle referentiedatum-afhankelijke informatie toegevoegd: bedrijfsparameters en eigenprestaties (geaggregeerde mpr-data, 305d-producties, lactatieproducties en vruchtbaarheidsdata). Hierbij werd uit de gepaste subdataset telkens de record opgehaald met de laatste datum vroeger dan of gelijk aan de referentiedatum. Voor de eigenprestaties werd dit voor alle lactaties gedaan die gestart werden vooraleer het beschouwde productiekengetal werd gerealiseerd, waarna de gegevens voor de verschillende lactaties kolomsgewijs werden geconcateneerd. Eigenprestatie-variabelen die onbekend waren omwille van het feit dat niet alle runderlevens perfect synchroon verlopen, kregen een nulwaarde toegekend. Dieren waarvoor de opgehaalde bedrijfsparameters meer dan een maand voor de referentiedatum werden berekend, werden niet weerhouden. Daarnaast werden voor de kengetallen van de groep *1^e lactatie* ook nog de eigenprestaties van de moeders aan de dataset toegevoegd. Dit gebeurde analoog aan de eigenprestaties van de dieren zelf, maar de referentiedatum werd verminderd met de tijdsperiode tussen de eerste kalving van de moeder en de geboortedatum van de beschouwde dochter. Dit werd gedaan omdat het op deze manier mogelijk werd gemaakt om gedetailleerd te bestuderen hoe een grotere hoeveelheid informatie over de moeder de resultaten van de machine learning modellen beïnvloedt. Dieren waarvoor de maternale eigenprestaties niet beschikbaar waren, of waarvan de

moeder haar leven had doorgebracht op een ander bedrijf, werden niet weerhouden in de inputdataset.

Tot slot werden nog een aantal extra variabelen toegevoegd aan de inputdatasets om (1) rekening te kunnen houden met seizoenseffecten en (2) de modellen duidelijke informatie te verschaffen over de levensfase van het dier op de referentiedatum: geboortemaand, maand van eerste inseminatie, maand van kalving (per lactatie), lactatie 1 gestart (0/1), lactatie 1 beëindigd (0/1), lactatie 2 gestart (0/1),... Dieren waarvoor een van deze variabelen op de referentiedatum niet gekend was, kregen hiervoor een nulwaarde toegekend.

Nadat de inputdataset op de juiste manier geassembleerd was, werden variabelen die enkel een discreet aantal waarden aannamen, zoals bijvoorbeeld de geboortemaand, omgezet in een passend aantal dummievariabelen. Het geheel aan dummievariabelen dat finaal in de inputdataset aanwezig was, liet de machine learning modellen toe om bij het ontbreken van gegevens over bepaalde lactaties, bepaalde mpr-metingen of bepaalde vruchtbaarheidsevents, zelf defaultwaarden te bepalen om mee te nemen in het model. Dit rechtvaardigt het toekennen van nulwaardes aan variabelen met betrekking tot (op de referentiedatum) ongekende lactatiegegevens of eigenprestaties.

Eenmaal voor een bepaald productiekengetal voor alle beschouwde referentiemomenten (zie §3.2) de inputdatasets waren opgesteld, werden alle dieren verwijderd die niet voor alle referentiemomenten werden weerhouden in de inputdataset. Bij de inputdatasets om de levensproductie te modelleren, werd deze post-assemblage selectiestap niet uitgevoerd, omdat anders het aantal records te klein dreigde te worden. Wel werden bij deze inputdatasets alle dieren verwijderd die op het beschouwde referentiemoment reeds waren afgevoerd, het heeft immers weinig zin om voor een dier dat op vier jaar leeftijd werd afgevoerd, vijf jaar na de geboorte de levensproductie te proberen voorspellen.

3.2 Beschouwde machine learning modellen

Om het bos door de bomen te blijven zien, worden in deze sectie drie begrippen geïntroduceerd die toelaten om elk van de opgestelde machine learning modellen eenvoudig te karakteriseren: de strategie, het referentiemoment en de gebruikte machine learning techniek.

De strategie beschrijft hoe het probleem wordt benaderd. Hierbij werden drie strategieën beschouwd: EVA, EVI en STIN. Bij EVA (Eén model Voor Allen) wordt één model

opgesteld op landelijk niveau, op basis van alle data aanwezig in de inputdataset. Bij EVI (Eén model Voor Ieder) daarentegen, wordt op bedrijfsniveau gewerkt: voor ieder bedrijf wordt een bedrijfsspecifiek machine learning model opgesteld. Voor de training en evaluatie van deze bedrijfsspecifieke modellen worden enkel en alleen maar de data van het beschouwde bedrijf gebruikt. Om bij EVI op degelijke wijze modellen te kunnen opstellen en evalueren, werden enkel maar bedrijven beschouwd waarvan minstens 25 records aanwezig waren in de inputdataset. Hierdoor is het totale aantal records dat in aanmerking komt om modellen te trainen en te evalueren bij EVI steeds iets kleiner dan bij EVA (Tabel 3.1). Als laatste strategie is er nog STIN (STackINg) die verder bouwt op EVA en EVI door op landelijk niveau één machine learning model te fitten op de originele inputdataset van EVI, uitgebreid met alle voorspellingen van de EVA- en EVI-modellen voor de betreffende records.

Tabel 3.1: Beschouwde referentiemomenten en karakteristieken van de inputdatasets. Met betrekking tot de karakteristieken van de inputdatasets, wordt het maximaal aantal variabelen, het aantal dieren en het aantal bedrijven weergegeven

kengetal	**referentie-momenten[a]**	**max # var.**	**EVA # dier**	**EVA # bedr.**	**EVI # dier**	**EVI # bedr.**
1e lactatie						
305d kg M		793	24 031	197	23 705	169
305d kg V		793	24 031	197	23 705	169
305d kg E		793	24 031	197	23 705	169
305d kg V+E	B, B1,..., B10	793	24 031	197	23 705	169
305d %V	I	793	24 031	197	23 705	169
305d %E	7K, 6K,..., K9	793	24 031	197	23 705	169
305d kg FCM		793	24 031	197	23 705	169
305d kg FPCM		793	24 031	197	23 705	169
tkt		793	21 579	196	21 242	166
x jaar lft.						
kg FPCM; 3 jaar	B, B1,..., B36	603	22 925	171	22 702	153
kg FPCM; 4 jaar	B, B1,..., B48	686	15 820	159	15 435	131
kg FPCM; 5 jaar	B, B1,..., B60	769	9606	143	9226	107
kg FPCM; 6 jaar	B, B1,..., B72	852	4998	120	4336	70
kg FPCM; 7 jaar	B, B1,..., B84	935	2159	102	1466	33
kg FPCM; 8 jaar	B, B1,..., B92	1018	783	70	258	7
leven						
kg FPCM[b]	B, B1,..., B108	1111				

[a] Bij STIN werd enkel maar referentiemoment B beschouwd
[b] De grootte van de inputdatasets met betrekking tot de levensproductie varieert in functie van de tijd: #dier EVA $\in [1423, 6302]$; #bedr. EVA $\in [71, 168]$; #dier EVI $\in [212, 5493]$; #bedr. EVI $\in [7, 104]$

Het referentiemoment beschrijft het tijdsaspect: op welk moment in het leven van het dier wordt het machine learning model opgesteld. Het referentiemoment is bijgevolg de bepalende factor voor de hoeveelheid informatie die beschikbaar is voor het opstellen van de machine learning modellen. Alle beschouwde referentiemomenten per productiekengetal werden gecodeerd aan de hand van een letter en eventueel een

getal (bv. B, B10, 7K) en worden weergegeven in Tabel 3.1. De letter geeft hierbij het ijkmoment weer: B voor geboorte, I voor de datum van eerste inseminatie en K voor de eerste kalving. Het cijfer drukt uit hoeveel maanden van 30.5 dagen er moeten worden opgeteld bij het ijkmoment (getal na letter) of moeten worden afgetrokken van het ijkmoment (getal voor letter) om de datum van het referentiemoment te bekomen. Indien er geen getal aanwezig is in de code, komt het referentiemoment overeen met het beschouwde ijkmoment. Zo betekent de code 7K: zeven maanden voor de eerste kalving. Betreffende de referentiemomenten berekend relatief tot over de datum van eerste kalving, moet opgemerkt worden dat er vanuit praktische overwegingen niet met referentiemomenten vroeger dan 7K werd gewerkt, drachtcontroles zijn immers maar mogelijk vanaf ongeveer 40 dagen dracht.

Het laatste 'begrip' is de machine learning techniek. In het kader van deze masterproef werden meer dan tien verschillende machine learning technieken beschouwd om productiekengetallen te voorspellen (Tabel 3.2) welke kunnen onderverdeeld worden in vier groepen. De eerste groep omvat machine learning technieken die nauw verwant zijn aan lineaire regressie: multiple lineaire regressie (de referentietechniek), ridge regressie en lasso regressie. De technieken in de tweede groep combineren principale componenten analyse (PCA) met een machine learning techniek nauw verwant aan lineaire regressie: (1) PCA + multiple lineaire regressie op de eerste n componenten die samen 99.99% van alle variantie verklaren (2) PCA + multiple lineaire regressie op de eerste 10 componenten (enkel EVI) (3) PCA+ multiple lineaire regressie op de eerste 100 componenten (EVA en STIN) (4) PCA + ridge regressie en (5) PCA + lasso regressie. Aangezien alle techieken die tot deze eerste twee groepen behoren reeds uitvoerig werden besproken in de literatuurstudie (§1.2.2) worden de achterliggende modelstructuren en modeleigenschappen hier niet opnieuw behandeld. De derde groep van beschouwde machine learning modellen bestaat uit technieken die gebaseerd zijn op regressiebomen: random forest regressie en een boosted regressiebomen. Deze modellen werden opgenomen omdat ze gekend staan voor hun brede toepasbaarheid en robuustheid, maar werden niet besproken in de literatuurstudie omwille van het feit dat ze amper worden gebruikt om fokwaarden te berekenen. Voor een bespreking van deze twee technieken gebaseerd op regressiebomen, wordt daarom verwezen naar Bijlage A. De vierde en laatste categorie omhelst support vector regressies, waarbij twee types kernels werden beschouwd: een radiale kernel en een polynomiale kernel. Ook de theorie achter support vector regressie werd uitvoerig besproken in de literatuurstudie en wordt daarom hier niet herhaald.

Tabel 3.2: Overzicht van de beschouwde machine learning technieken

afkorting	**techniek**
linreg	multiple lineaire regressie
ridge	ridge regressie
lasso	lasso regressie
PCA + linreg	principale componenten regressie (multiple lineaire regressie)
PCA_10	principale componenten regressie (multiple lineaire regressie op de eerste 10 componenten)
PCA_100	principale componenten regressie (multiple lineaire regressie op de eerste 100 componenten)
PCA + ridge	principale componenten regressie (ridge regressie)
PCA + lasso	principale componenten regressie (lasso regressie)
random forest	random forest regressie
boosting tree	boosted regressiebomen
SVR_radial	support vector regressie (radiale kernel)
SVR_poly	support vector regressie (polynomiale kernel)

3.3 Opstellen en evaluatie van de machine learning modellen

Hierboven werd besproken welke modeltechnieken beschouwd werden, welke strategieën werden toegepast en voor welke referentiemomenten er modellen werden opgesteld. In deze sectie wordt besproken hoe de machine learning modellen effectief opgesteld en geëvalueerd werden.

Bij EVA werden, na het inlezen van de inputdataset, eerst alle variabelen verwijderd waarvoor minder dan 2% van de records een waarde had verschillend van de modus. Vervolgens werden alle variabelen en de te modelleren productiekengetallen gestandaardiseerd ($\mu = 0$, $\sigma^2 = 1$). Na het standaardiseren van de dataset, werden de dieren in vijf gelijke *folds* ingedeeld, waarbij de eerste *fold* als testdataset werd gebruikt en de resterende vier *folds* samengevoegd werden tot een trainingsdataset. Om consistente resultaten te garanderen, werden per productiekengetal slechts eenmaal willekeurig *folds* opgesteld (voor referentiemoment B) waarna deze *fold*-indeling vervolgens voor alle beschouwde referentiemomenten werd gehanteerd.

Eenmaal deze voorbereidende handelingen uitgevoerd waren, werden de machine learning modellen (Tabel 3.2) een na een getraind, waarbij gebruik werd gemaakt

van functies uit het Python scikit-learn package. Voor het tunen van eventuele hyperparameters werd steeds een voldoende brede range aan mogelijke waarden in beschouwing genomen (vb. 1, 2, 3, 4 en 5 voor de graad d van de polynomiale kernel in vergelijking 1.22) waarbij via een *gridsearch* met cross-validatie uit de beschouwde range de meest optimale waarde werd gekozen. De enige hyperparameter die op voorhand werd vastgelegd, was het aantal regressiebomen in het random forest, dat werd vastgelegd op 500.

Nadat alle machine learning modellen opgesteld waren, werden de productiekengetallen van de dieren in de testdataset voorspeld met behulp van de opgestelde machine learning modellen. Deze voorspellingen werden vervolgens opnieuw getransformeerd naar de oorspronkelijke distributie van het productiekengetal (oorspronkelijke μ en σ^2) en zorgvuldig opgeslagen om later de modellen te kunnen evalueren. Indien de testdataset geen 1000 records bevatte, werden alle modellen opnieuw opgesteld, maar deze keer met de tweede *fold* als testdataset. Op deze manier werden de verschillende *folds* een na een overlopen, tot er een testdataset kon worden samengesteld met meer dan 1000 records, of alle vijf de *folds* als testdataset hadden gefungeerd. Bij het modelleren van het kengetal levensproductie (kg FPCM) werd afgeweken van deze evaluatieregel en werden steeds alle vijf de *folds* eenmaal als testdataset gebruikt.

De methodologie die gebruikt werd bij EVI is volledig analoog aan de methodologie die gebruikt werd bij EVA, met dat verschil dat de hele procedure op bedrijfsniveau werd uitgevoerd. Dit houdt in dat via een for-lus uit de inputdataset telkens de data van een enkel bedrijf werden geïsoleerd, waarna de hierboven beschreven procedure met de bedrijfsspecifieke data werd doorlopen. Doordat het maximaal aantal records per bedrijf 1001 was, wordt hierbij voor alle bedrijven de procedure steeds vijfmaal doorlopen, iedere keer met een andere *fold* als testdataset. Aangezien bij EVI ook de normalisatie op bedrijfsniveau werd uitgevoerd, is de terug-transformatie van de modelvoorspellingen naar de oorspronkelijke distributie essentieel om de EVI-modellen op landelijk niveau te kunnen evalueren.

Bij STIN werd exact dezelfde procedure gevolgd als bij EVA, maar moest eerst de inputdataset van EVI uitgebreid worden met alle voorspellingen van de EVA- en EVI-modellen. Een aandachtspunt hierbij, was dat er geen informatie over de werkelijke kengetallen mocht worden ingebracht in de STIN-inputdataset. Daarom werd de EVA procedure vijfmaal doorlopen bij het opstellen van de inputdataset voor STIN, zodat ook bij EVA voor ieder dier voorspellingen beschikbaar waren zonder risico op overfitting. Omdat hierdoor de benodigde rekentijd sterk toenam, werd voor STIN, zoals eerder reeds vermeld, enkel referentiemoment B (geboorte) beschouwd. Aangezien bij het opstellen van de STIN-inputdataset voor alle dieren voorspellingen zonder over-

fitting beschikbaar kwamen voor de EVA- en EVI-modellen, was het een kleine moeite om ook voor STIN steeds alle vijf de *folds* eenmaal als testdataset te gebruiken, zodat bij het evalueren van de verschillende machine learning technieken op bedrijfsniveau voor referentiemoment B, steeds ruim voldoende records per bedrijf beschikbaar waren om deze evaluatie op een degelijke manier uit te kunnen voeren.

Tijdens de effectieve evaluatie van de modellen werden verschillende statistieken berekend op basis van de voorspellingen die werden gemaakt voor de dieren in de testdataset. De eerste statistiek die werd berekend is R^2 op landelijk niveau ($R^2_{landelijk}$). Bij EVI werden hiervoor de voorspellingen van de verschillende bedrijven samengebracht in een vector $\hat{\boldsymbol{y}}$ en werden alle werkelijke productiekengetallen samengebracht in een vector $\boldsymbol{y}$, waarna op basis van deze twee vectoren $R^2_{landelijk}$ werd berekend. Daarnaast werd ook de vierkantswortel van de mean squared error ($\sqrt{\text{MSE}}$; mean squared error = gemiddelde kwadratische fout) op landelijk niveau berekend, welke een maat is voor de standaardafwijking van predictiefouten. Als laatste statistiek werden R^2-waarden berekend op bedrijfsniveau ($R^2_{bedrijf}$) hierbij werden enkel bedrijven beschouwd waarvoor minstens 25 dieren in de testdataset aanwezig waren.

3.4 Resultaten en bespreking

3.4.1 Equivalente techieken en overfitting

De resultaten van de evaluatie op landelijke niveau ($R^2_{landelijk}$ en $\sqrt{\text{MSE}}$) van de opgestelde EVA- en EVI-machine learning modellen, worden weergegeven in de figuren die zijn opgenomen in Bijlage B. Voor alle productiekengetallen geldt hierbij dat $-R^2 \sim (\sqrt{\text{MSE}})^2$. Enkel de levensproductie vormt hierop een uitzondering doordat voor dit productiekengetal de variantie van de gemodelleerde distributie niet constant is doorheen de tijd als gevolg van het feit dat dieren die op het beschouwde referentiemoment reeds werden afgevoerd, niet werden opgenomen in de inputdataset. Voor alle productiekengetallen, met uitzondering van de levensproductie, geven de figuren voor $R^2_{landelijk}$ en $\sqrt{\text{MSE}}$ dus eigenlijk dezelfde info weer. Toch werd er voor gekozen om steeds beide statistieken weer te geven, aangezien enerzijds de waarden voor $R^2_{landelijk}$ eenvoudig te vergelijken zijn tussen de verschillende productiekengetallen, terwijl anderzijds $\sqrt{\text{MSE}}$ dezelfde eenheid heeft als het beschouwde productiekengetal, wat de interpretatie van de modelperformantie vereenvoudigt. Aanvullend op de grafieken in Bijlage B, worden de exacte waarden voor $R^2_{landelijk}$ voor referentiemoment B (geboorte) van het beste model per strategie daarnaast weergegeven in Tabel 3.3.

Bij de analyse van de resultaten was een van de eerste bevindingen dat PCA + ridge voor alle beschouwde kengetallen op alle beschouwde referentiemomenten exact dezelfde modelperformanties realiseerde als ridge regressie. Bij nader onderzoek van de modelvoorspellingen zelf, bleken ook deze exact overeen te komen. Principale componenten analyse uitvoeren vooraleer ridge regressie toe te passen biedt dus geen enkele meerwaarde en vraagt alleen maar meer rekentijd. PCA + ridge wordt dan ook niet verder besproken en wordt ook niet weergegeven in de figuren in Bijlage B.

Enkel de modellen die op alle referentiemomenten opgesteld konden worden en niet te veel overfitting vertoonden, worden weergegeven in de figuren in Bijlage B. Praktisch betekent dit bij EVA dat voor de meeste productiekengetallen alle beschouwde machine learning technieken worden weergegeven. Enkel linreg en PCA + linreg worden bij kg FPCM op 7/8 jaar leeftijd en de levensproductie, niet altijd weergegeven (figuren B.27 - B.32). Door een lage verhouding waarnemingen/parameters bij deze productiekengetallen, waren deze technieken immers sterk onderhevig aan overfitting (met regelmatig $R^2_{landelijk} \leq 0$ tot gevolg) of konden zij zelfs helemaal niet toegepast worden omdat het aantal modelparameters groter was dan het aantal waarnemingen. PCA + linreg heeft iets minder vlug last van overfitting dan linreg, maar is er niet ongevoelig voor, zo kan in Figuur B.27 gezien worden dat ook deze methode bij een beperkt aantal records snel wisselvallige resultaten toont voor $R^2_{landelijk}$ als gevolg van overfitting.

Bij EVA is het aantal machine learning modellen dat niet opgesteld kon worden of slechte resultaten opleverde als gevolg van overfitting dus relatief beperkt. Dit ligt anders voor EVI, waar linreg en PCA + linreg nooit worden weergegeven in de figuren in Bijlage B aangezien het aantal te bepalen modelparameters nagenoeg altijd groter was dan het aantal waarnemingen dat per bedrijf beschikbaar was. Het aantal te bepalen modelparameters werd bij EVI weliswaar sterk gereduceerd door de stap in het algoritme waarbij alle variabelen worden verwijderd waarvoor minder dan 2% van de records een waarde vertoont verschillend van de modus, maar dit was nooit voldoende om voor alle bedrijven met meer dan 25 beschikbare records linreg of PCA + linreg te kunnen toepassen.

Bij EVI blijkt ook PCA_10 gevoelig aan overfitting: met betrekking tot de kg FPCM-productie op een leeftijd van x-jaar begint de curve voor $R^2_{landelijk}$ abnormaal grote fluctuaties te vertonen vanaf x=5. Naarmate x verder toeneemt daalt het aantal waarnemingen per bedrijf verder, terwijl het maximaal aantal variabelen in de dataset steeds maar toeneemt. Hierdoor fluctueert de curve voor $R^2_{landelijk}$ in Figuur B.57 (x=6) en B.59 (x=7) meer dan in Figuur B.55 (x=5). Voor x = 8 is de overfitting zelfs

zo erg dat de $R^2_{landelijk}$ regelmatig negatief is, waardoor de curve voor PCA_10 in Figuur B.61 niet meer wordt weergegeven. Analoog treedt er bij de levensproductie vanaf ongeveer 60 maanden leeftijd sterke overfitting bij PCA_10 op, waardoor PCA_10 ook in Figuur B.63 niet wordt weergegeven.

De hierboven beschreven waarnemingen wijzen onmiddellijk op een eerste voordeel van machine learning technieken tegenover klassieke lineaire regressie, namelijk dat (bepaalde) machine learning technieken veel minder/amper gevoelig zijn voor overfitting en daardoor ook bij een beperkt aantal waarnemingen kunnen toegepast worden. Machine learning technieken die in alle EVI-figuren (B.33 - B.64) worden weergegeven, hebben bijvoorbeeld met 20 waarnemingen ($25 \cdot 0.8 = 20$) voldoende om modellen zonder ernstige overfitting op te stellen. Een alternatief om toch lineaire regressie te kunnen toepassen is modelselectie, waarbij, om overfitting te vermijden, meestal enkel de meest significante variabelen worden weerhouden. Dit impliceert echter dat de informatie die in de overige variabelen aanwezig is, niet wordt gebruikt, wat vaak ten koste gaat van de modelperformantie. In de figuren in Bijlage B kan zo bijvoorbeeld gezien worden dat PCA_100 en PCA_10 opvallend vaak beduidend lagere performanties realiseren dan alle andere beschouwde machine learning technieken, dit als gevolg van het feit dat de informatie die vervat zit in de laatste principale componenten, niet gebruikt wordt. Indien men principale componenten regressie wil toepassen in het kader van het voorspellen van productiekengetallen van melkvee, is het daarom aangeraden om lasso (of ridge) regressie toe te passen op de bekomen componenten en niet simpelweg multiple lineaire regressie uit te voeren op de eerste 10/100 componenten.

3.4.2 Performantie op referentiemoment B (geboorte)

EVA

Wat betreft EVA, kan in Tabel 3.3 gezien worden dat voor alle 8 de beschouwde productiekengetallen met betrekking tot de 1^e lactatie, SVR_radial de beste resultaten oplevert op referentiemoment B (geboorte). In de figuren in Bijlage B kan daarnaast gezien worden dat, voor de productiekengetallen van de groep 1^e lactatie, tree boosting steeds de op een na beste techniek is, waarna alle andere machine learning technieken op korte afstand volgen. Enkel PCA_100 levert, om eerder uiteengezette redenen, steeds een beduidend lager modelperformantie op. Deze waarneming kan verklaard worden aan de hand van de modelstructuur van de verschillende machine learning technieken. Zoals in de literatuurstudie uiteengezet, zijn machine learning technieken zoals ridge regressie, lasso regressie en principale componenten regressie

Tabel 3.3: $R^2_{landelijk}$ voor de beste machine learning modellen per strategie, geëvalueerd op referentiemoment B (geboorte). Machine learning technieken die een $R^2_{landelijk}$ realiseerden die minder dan 0.01 verschilde van deze van het beste model, worden weergegeven met een superscript

	EVA		EVI		STIN	
kengetal	**techniek**	**R^2**	**techniek**	**R^2**	**techniek**	**R^2**
1e lactatie						
305d kg M	SVR_radial	0.55	lasso	0.48	linreg[bg]	0.57
305d kg V	SVR_radial	0.50	lasso	0.45	linreg[bcg]	0.54
305d kg E	SVR_radial[i]	0.50	random forest[c]	0.45	PCA + lasso[ab]	0.54
305d kg V+E	SVR_radial	0.49	random forest[bc]	0.44	PCA + lasso[abc]	0.53
305d %V	SVR_radial[i]	0.69	lasso	0.63	linreg[bcg]	0.71
305d %E	SVR_radial	0.64	lasso	0.50	linreg[bg]	0.66
305d kg FCM	SVR_radial	0.49	lasso[bh]	0.43	linreg[bg]	0.54
305d kg FPCM	SVR_radial	0.50	random forest[bc]	0.44	linreg[bg]	0.54
tkt	linreg[bcdg]	0.08	ridge[cg]	0.03	linreg[bcg]	0.12
x jaar lft.						
kg FPCM; 3 jaar	SVR_radial	0.43	ridge[h]	0.40	linreg[bcg]	0.47
kg FPCM; 4 jaar	tree boosting[abcgk]	0.50	random forest[bc]	0.47	linreg[bg]	0.56
kg FPCM; 5 jaar	SVR_radial[abcg]	0.55	random forest[c]	0.48	ridge[acg]	0.60
kg FPCM; 6 jaar	SVR_radial[c]	0.53	random forest	0.45	linreg[bcg]	0.61
kg FPCM; 7 jaar	lasso[bgk]	0.55	random forest	0.50	PCA + lasso[abc]	0.68
kg FPCM; 8 jaar	lasso[bm]	0.59	tree boosting[bghk]	0.50	lasso[gh]	0.60
leven						
kg FPCM	PCA + lasso	0.30	random forest	0.22	PCA + lasso[abc]	0.47

[a] linreg; [b] ridge; [c] lasso; [d] PCA + linreg; [e] PCA_10; [f] PCA_100; [g] PCA + lasso; [h] random forest; [i] boosting tree; [k] SVR_radial; [m] SVR_poly

immers ontwikkeld om modellen met dezelfde modelstructuur als de klassieke multiple lineaire regressie op te stellen, maar waarbij deze machine learning technieken minder gevoelig zijn aan overfitting en multicollineariteit. Voor de productiekengetallen van de groep 1e lactatie spelen overfitting en multicollineariteit bij EVA echter maar een beperkte rol, er zijn immers steeds ruim voldoende waarnemingen en bij het opstellen van de inputdataset werd multicollineariteit zo veel mogelijk vermeden, onder andere door het gebruik van principale componenten analyse bij het berekenen van de variabelen met betrekking tot het fokwaarde × bedrijf interactie-effect. Als gevolg hiervan leveren ridge, lasso, PCA + linreg en PCA + lasso steeds gelijkaardige resultaten op als linreg, waardoor deze machine learning technieken nagenoeg op elkaar liggen in figuren B.1 - B.16

SVR_poly vervoegt deze 'bundel' van machine learning technieken als gevolg van het feit dat de graad d van de polynomiale kernel bij het opstellen van de machine learning modellen steeds de waarde 1 toegekend kreeg, op enkele sporadische uitzonderingen na. Aangezien een polynomiale kernel met graad 1, op een constante na, gelijk is aan een lineaire kernel (zie vgl. (1.22) en (1.20)) impliceert dit immers dat de opgestelde support vector regressie geen rekening kan houden met interactie-effecten tussen de verschillende variabelen en aldus gelijkaardige performanties zal

realiseren als de 'bundel' van machine learning technieken die rond linreg gesitueerd is. Het feit dat SVR_radial en tree boosting wél interactie-effecten tussen de verschillende variabelen in rekening kunnen brengen, verklaart dan ook meteen waarom het juist deze machine learning technieken zijn die de beste performanties realiseren, al is het verschil met de lineaire technieken steeds relatief beperkt en nooit groter dan 0.04. Ondanks dat random forest ook in staat is om interactie-effecten tussen de verschillende variabelen in de dataset in rekening te brengen, realiseert deze modeltechniek geen hogere $R^2_{landelijk}$ dan linreg, ridge, lasso,... Mogelijks is dit te wijten aan de zeer hoge robuustheid van deze machine learning techniek, die bij een voldoende groot aantal waarnemingen eerder leidt tot een iets lagere dan een iets hogere performantie, zeker als de relatie tussen de inputvariabelen en het te voorspellen productiekengetal vrij lineair is, wat bij alle productiekengetallen met betrekking tot de 1^e lactatie het geval lijkt te zijn.

Wordt de grootteorde van $R^2_{landelijk}$ in Tabel 3.3 voor EVA bekeken, dan kan gezien worden dat SVR_radial er reeds bij de geboorte van een dier in slaagt om voor alle beschouwde productiekengetallen met betrekking tot de eerste lactatie minimaal ±50% van variatie op landelijk niveau te verklaren. Aangezien deze waarden voor $R^2_{landelijk}$ steeds hoger liggen dan de corresponderende heritabiliteiten die door CRV (2020a) worden gerapporteerd (0.48 voor kg M, 0.48 voor kg V en 0.39 voor kg E) kan besloten worden dat, in vergelijking met lineaire regressie op de fokwaarden, de opgestelde machine learning modellen zeker een meerwaarde bieden om productiekengetallen van melkvee te voorspellen. Gegeven het feit dat ook lineaire regressie relatief goede waarden voor $R^2_{landelijk}$ realiseert, speelt hierbij de samenstelling van de inputdataset hoogstwaarschijnlijk een grotere rol dan het gebruik van machine learning technieken op zichzelf. De hoge waarden voor $R^2_{landelijk}$ die worden gerealiseerd voor %V en %E (305d, 1^e lactatie) zijn hierbij zeer opvallend, zeker omdat de fokwaarden voor deze productiekengetallen ontbreken in de door CRV aangeleverde dataset. SVR_radial slaagt er hierbij reeds bij de geboorte in om respectievelijk 69% en 64% van de (landelijke) variatie in vet- en eiwitgehalte van de melk te verklaren, wat verrassend veel is.

Wordt er bij EVA gekeken naar de performantie van de verschillende machine learning technieken op referentiemoment B voor wat betreft de tkt tussen de eerste en de tweede pariteit (Figuur B.17) dan kan gezien worden dat de prestaties van de opgestelde machine learning modellen niet bepaald spectaculair zijn. Klassieke multiple lineaire regressie levert hierbij op referentiemoment B met een $R^2_{landelijk}$ van 0.08 de 'beste' resultaten op, maar verschillende andere machine learning technieken (ridge, lasso, PCA + linreg en PCA + lasso) volgen op korte afstand. Ook van de hogere performantie die SVR_radial en tree boosting realiseerden bij de productiekengetallen

van de groep 1^{e} lactatie, blijft bij het modelleren van de tkt niet veel meer over, de support vector regressies (SVR_radial en SVR_poly) scoren zelfs het slechtst van alle beschouwde machine learning technieken.

Wordt in Tabel 3.3 bij EVA gekeken naar de productiekengetallen van de groep x jaar leeftijd, dan kan gezien worden dat binnen deze groep de dominantie van SVR_radial niet meer in dezelfde mate aanwezig is in vergelijking met de productiekengetallen van de groep 1^{e} lactatie. Wordt in figuren B.19 - B.26 gekeken hoe de andere machine learning technieken presteren voor referentiemoment B (geboorte), dan kan gezien worden dat de verschillende beschouwde machine learning technieken, steeds heel gelijkaardige performanties opleveren, wat zich uit in het feit dat in Tabel 3.3 tot vijf verschillende modeltechnieken worden gerapporteerd die een $R^2_{landelijk}$ realiseren die minder dan 0.01 verschilt van de $R^2_{landelijk}$ van de beste beschouwde machine learning techniek. Voor de kg FPCM-productie op 7 en 8 jaar leeftijd liggen de waarden voor $R^2_{landelijk}$ iets meer uit elkaar, maar niet in die mate dat een machine learning als de beste kan bestempeld worden.

Wordt gekeken naar de nominale waarde van $R^2_{landelijk}$ van de best presterende machine learning modellen op referentiemoment B voor de kg FPCM-productie op x jaar leeftijd, dan kan gezien worden dat deze in dezelfde grootteorde liggen als deze van de productiekengetallen van de groep 1^{e} lactatie. Wordt in figuren B.20 - B.30 gekeken naar de grootte van $\sqrt{\text{MSE}}$, dan kan wel gezien worden dat deze stelselmatig toeneemt van ongeveer 1450 kg FPCM op 3 jaar leeftijd naar ongeveer 4500 kg FPCM op 8 jaar leeftijd. Verondersteld dat de residuen normaal verdeeld zijn, betekent dit dat voor 95% van de dieren de voorspelde productie op 3 jaar en 8 jaar leeftijd respectievelijk maximaal 2900 en 9000 kg FPCM afwijkt van de werkelijke kg FPCM-productie. Gegeven een gemiddelde 305d-productie van de Nederlandse melkkoeien in 2019 van 9837 kg FPCM (CRV, 2020b) is het niet moeilijk om in te zien dat de voorspellingen dus alles behalve accuraat zijn, zelfs al verklaren de opgestelde machine learning modellen tot meer dan 50% van de landelijke variatie in deze productiekengetallen. Wel is het zo dat de standaardafwijking van de kg FPCM productie op x jaar leeftijd relatief gezien nog sneller stijgt dan $\sqrt{\text{MSE}}$ naarmate x toeneemt, wat zich uit in een stijgende trend van de $R^2_{landelijk}$ in Tabel 3.3 naarmate de beschouwde leeftijd x toeneemt, en dit zowel voor EVA, EVI als STIN. Dit lijkt op het eerste zicht misschien niet logisch, maar als er van uit gegaan wordt dat de verklaarbare effecten op de melkproductie positief gecorreleerd zijn over verschillende lactaties en de onverklaarbare effecten onafhankelijk zijn tussen de verschillende lactaties, is het niet moeilijk om wiskundig aan te tonen dat het aandeel van de verklaarbare variantie in de totale variante toeneemt naarmate het aantal beschouwde lactaties/de beschouwde leeftijd x toeneemt.

Als laatste productiekengetal dat binnen de EVA-strategie kan besproken worden is er de levensproductie, waarbij PCA + lasso op referentiemoment B een $R^2_{landelijk}$ realiseert van 0.30, op korte afstand gevolgd door ridge en lasso (Figuur B.31) SVR_radial, SVR_poly, tree boosting en random forest volgen op iets groter afstand en realiseren waarden voor $R^2_{landelijk}$ rond 0.25. De maximaal gerealiseerde $R^2_{landelijk}$ mag dan wel net geen 4 keer zo groot zijn als deze die werd bekomen voor de tkt tussen de eerste en de tweede pariteit, $R^2_{landelijk}$ blijft laag en de voorspellingen zijn bijgevolg helemaal niet nauwkeurig: op referentiemoment B ligt $\sqrt{\text{MSE}}$ bij de meeste technieken rond 16000 kg FPCM, wat overeenstemt met 1.65 keer het gemiddelde rollend jaargemiddelde van de Nederlandse koeien in 2019 (CRV, 2020b).

EVI

Bij EVA was SVR_radial vaak de beste machine learning techniek omdat deze interactie-effecten tussen de variabelen kon meenemen. Bij EVI daarentegen domineren lasso regressie en random forest de resultaten in Tabel 3.3, dit omdat deze technieken goed om kunnen gaan met het beperkte aantal waarnemingen dat per bedrijf beschikbaar is. Bij de productiekenmerken met betrekking tot de eerste lactatie, is hierbij het verschil tussen de beste en de slechtste machine learning techniek regelmatig groter is dan 0.10 Dit is vooral opvallend in figuren B.42 en B.44, waar lasso regressie voor %V en %E een $R^2_{landelijk}$ realiseert van respectievelijk 0.63 en 0.50 en het verschil met de eerstvolgende techniek telkens ongeveer 0.10 bedraagt. Dit betekent dat lasso regressie er voor deze kengetallen in slaagt om 10% meer van de landelijke variantie te verklaren dan de eerstvolgende machine learning techniek.

Wat betreft de nominale waarden van $R^2_{landelijk}$ voor de verschillende productiekengetallen, kan in Tabel 3.3 gezien worden dat deze zonder uitzondering kleiner zijn dan deze voor EVA, waarbij de waarden voor $R^2_{landelijk}$ bij EVI 0.03-0.14 lager liggen dan de corresponderende waarden voor $R^2_{landelijk}$ bij EVA. Deze lagere waarden voor $R^2_{landelijk}$ bij EVI zijn het gevolg van een trade-off tussen enerzijds de mogelijkheid van de modellen om zich volledig te focussen op de intra-bedrijfsvariantie en anderzijds de beperkte hoeveelheid records die per bedrijf beschikbaar is. Bij EVI weegt de beperkte databeschikbaarheid duidelijk zwaarder door dan de mogelijkheid van de modellen om zich volledig te focussen op de intra-bedrijfsvariantie, wat leidt tot lagere waarden voor $R^2_{landelijk}$ in vergelijking met EVA.

STIN

Wordt tot slot in Tabel 3.3 gekeken naar de resultaten voor STIN, dan valt onmiddellijk op dat de waarden voor $R^2_{landelijk}$ die gerapporteerd worden steeds hoger zijn dan deze voor EVA en EVI. Bovendien zitten 4 machine learning technieken elkaar steeds op de hielen in de strijd om als beste techniek vermeld te staan in Tabel 3.3: linreg, ridge, lasso en PCA + lasso. Random forest, tree boosting, SVR_radial en SVR_poly lijken in het kader van STIN minder geschikte machine learning technieken, die de meerwaarde van stacking niet of amper kunnen uitbuiten. Voor kg FPCM op 5 jaar leeftijd bijvoorbeeld, realiseren linreg, ridge, lasso en PCA + lasso allemaal een $R^2_{landelijk}$ tussen 0.59 en 0.60, terwijl random forest, tree boosting, SVR_radial en SVR_poly waarden voor $R^2_{landelijk}$ realiseren van maximaal 0.54, wat in lijn ligt met de waarden die voor $R^2_{landelijk}$ werden bekomen bij EVA.

Vooral bij de levensproductie is de toename van R^2 spectaculair: door de verschillende EVA- en EVI-modellen te combineren stijgt $R^2_{landelijk}$ met maar liefst 0.17 tot 0.47, wat in dezelfde grootte-orde ligt als de performanties die bij EVA werden bekomen voor het voorspellen van de 305d-productie in de eerste lactatie, uitgedrukt in kg V, kg E, kg V+E, kg FCM of kg FPCM. Matig presterende EVA- en EVI-modellen zijn echter geen voorwaarde opdat STIN een meerwaarde zou kunnen bieden, zo realiseert STIN voor 305d %V tijdens de eerste lactatie een R^2 van 0.71, wat nog eens een stijging van 0.02 inhoudt in vergelijking met de reeds hoge waarde voor $R^2_{landelijk}$ die bekomen werd met SVR_radial bij EVA.

Stacking biedt dus overduidelijk een meerwaarde bij het voorspellen van productiekengetallen. Dit komt hoogstwaarschijnlijk doordat de verschillende beschouwde technieken elk hun sterke en zwakke punten hebben, waarbij door het combineren van de verschillende technieken via stacking de sterktes van de verschillende technieken elkaar kunnen aanvullen en de zwaktes kunnen gecompenseerd worden door sterktes van andere machine learning technieken. Zo is SVR_radial bij voldoende databeschikbaarheid (EVA) een ideale techniek om interactie-effecten tussen de verschillende variabelen in de dataset in rekening te brengen, terwijl lasso en random forest goed zijn in het modelleren van de variabiliteit binnen een enkel bedrijf met een beperkte databeschikbaarheid (EVI). Voor stacking blijkt bovendien ook geen ingewikkeld model nodig te zijn, vaak levert eenvoudige multiple lineaire regressie zelfs de beste resultaten op. Machine learning technieken bieden dus een overduidelijke meerwaarde in het kader van het voorspellen van productiekengetallen, waarbij het niet één individueel model is dat multiple lineaire regressie op landelijk niveau overstijgt, maar het vooral de combinatie van de verschillende machine learning technie-

ken via stacking is die leidt tot een duidelijke stijging van de performantie waarmee productiekengetallen van melkvee kunnen voorspeld worden.

3.4.3 Performantie doorheen de tijd

Naast de performantie van de verschillende machine learning technieken op referentiemoment B (geboorte), werd bij EVA en EVI ook bestudeerd in hoeverre de beschouwde machine learning technieken in staat zijn om bijkomende informatie te gebruiken om de modelperformantie te verhogen.

EVA

Wat betreft de productiekengetallen met betrekking tot de eerste lactatie bij EVA (Figuur B.1 - B.16) vertonen alle productiekengetallen een analoog patroon. Vertrekkende van het performantie-niveau dat de verschillende technieken behalen op referentiemoment B (geboorte) verlopen de curves voor $R^2_{landelijk}$ en $\sqrt{\text{MSE}}$ vrijwel horizontaal tot en met referentiemoment I (eerste inseminatie). Geen enkele van de machine learning modellen is dus in staat om de extra informatie die beschikbaar komt (maternale eigenprestaties en leeftijd bij eerste inseminatie) om te zetten in een hogere modelperformantie. Hoogstwaarschijnlijk komt dit doordat de leeftijd van eerste inseminatie weinig tot geen invloed heeft op de productiekengetallen tijdens de eerste lactatie. Waarom het beschikbaar komen van de maternale eigenprestaties niet leidt tot een hogere modelperformantie wordt besproken in §3.4.4

Vanaf referentiemoment 7K stijgt echter voor alle productiekengetallen, met uitzondering van 305d %V en %E tijdens de eerste lactatie, de performantie van de beste machine learning techniek. Deze stijging is duidelijk, maar relatief beperkt, wat aangeeft dat de leeftijd bij eerste kalving (die vanaf referentiemoment 7K gekend is) een bescheiden invloed heeft op de productiekengetallen (met uitzondering van %V en %E) die zullen worden gerealiseerd tijdens de eerste lactatie (zie §3.4.7). De meeste machine learning technieken volgen deze stijging in performantie van het best presterende model, enkel random forest slaagt er niet in om deze extra bron van nuttige informatie te vertalen in een hogere $R^2_{landelijk}$. Vanaf referentiemoment 7K verlopen de curves terug horizontaal, tot referentiemoment K,vanaf wanneer de performantie stijgt als gevolg van het beschikbaar komen van eigenprestaties van de dieren.

Deze performantiestijging verloopt initieel zeer snel, zo stijgt $R^2_{landelijk}$ voor SVR_radial in Figuur B.1 van 0.57 bij kalving op 3 maanden tijd naar 0.84 Aangezien alle beschouwde machine learning methoden deze snelle stijging vertonen, ligt de oorzaak van deze snelle stijging niet bij het gebruiken van machine learning technieken, maar

wel bij het juist aanreiken van nuttige informatie, wat wijst op het belang van een doordachte primaire dataverwerking (§2.3). Deze stijging vlakt enigzins af naarmate de lactatie vordert, wat aangeeft dat de producties die een bepaald dier realiseert tijdens de eerste maanden van haar lactatie, ook veel informatie bevatten over hoe het verdere verloop zal zijn van de lactatie, wat niet geheel onlogisch is. Opvallend hierbij is dat deze afvlakking meer uitgesproken is voor random forest, wat verklaard kan worden door de modelstructuur die wordt gebruikt bij deze techniek (§A) die maar moeilijk toelaat om waarden voor $R^2_{landelijk}$ te realiseren die dicht bij 1 liggen.

Wat betreft de tkt tussen pariteit 1 en 2, is in figuren B.17 en B.18 te zien dat de performantie van alle beschouwde machine learning technieken begint te stijgen vanaf referentiemoment K2, wat verklaard kan worden omdat vanaf dat refentiemoment informatie beschikbaar komt met betrekking tot de inseminaties van de dieren tijdens de eerste lactatie. Opvallend in Figuur B.17 is de terugval die alle machine learning technieken, met uitzondering van random forest en boosting tree, vertonen op referentiemoment K6. Op dit moment is er nog een beperkt aantal dieren dat nog steeds niet voor de eerste keer geïnsemineerd is (maar dat wel een tkt zal realiseren) en waarvoor de meeste machine learning modellen onrealistische voorspellingen maken als gevolg van het feit dat de modelcoëfficiënten niet geoptimaliseerd zijn om de tkt van dit type dieren te voorspellen. Enkel regressieboom-gebaseerde machine learning technieken produceren door hun robuustheid geen onrealistische voorspellingen voor deze dieren, wat de performantie van deze modellen op referentiemomenten K6-8 ten goede komt.

Ook de performanties van de EVA-modellen voor de kg FPCM-productie op een leeftijd van x jaar vertonen een analoog patroon doorheen de tijd (figuren B.19 - B.30), maar bij deze productiekengetallen begint $R^2_{landelijk}$ te stijgen tussen 12 maanden en 24 maanden leeftijd, afhankelijk van welke leeftijd x men beschouwt. Voor lage waarden van x ligt het buigpunt eerder rond 12 maanden, het moment wanneer de eerste pinken geïnsemineerd worden. Dit is niet geheel onlogisch, aangezien de leeftijd waarop de dieren afkalven, een grote invloed heeft op de tijdsperiode gedurende dewelke de dieren melk kunnen produceren vooraleer ze de leeftijd van 3/4/5 jaar bereiken. Voor grote waarden van x (7 of 8) is het vooral het productieniveau dat een grote rol speelt bij de kg FPCM-productie die zal worden gerealiseerd, waardoor het buigpunt dan ook eerder op 24 maanden leeftijd ligt, het moment dat van de eerste dieren mpr-gegevens beschikbaar komen. Omdat de leeftijd bij eerste kalving zo belangrijk is voor de gerealiseerde productie op 3/4/5 jaar leeftijd, vertonen, analoog aan hoe dit voor de tkt was, alle modellen met uitzondering van random forest en tree boosting een daling in performantie als de dieren een leeftijd hebben van 19 maanden. Op dat moment zijn immers de meeste dieren reeds drachtig, met uitzondering van een

aantal uitschieters die nog drachtig zullen worden, maar het op 19 maanden leeftijd nog steeds niet zijn. Aangezien het belang van de leeftijd bij eerste afkalving daalt naarmate de waarde van x groter wordt, daalt ook de grootteorde van de performantiedip: deze is het grootste bij x = 3, kleiner bij x = 4 en amper nog zichtbaar bij x = 5.

Met betrekking tot de levensproductie is op Figuur B.32 is te zien dat, vertrekkende van de initiële $\sqrt{\text{MSE}}$ van ongeveer 16000 kg FPCM, $\sqrt{\text{MSE}}$ vanaf een leeftijd van ongeveer 24 maanden langzaam zakt. $\sqrt{\text{MSE}}$ blijft echter steeds hoog, zo is $\sqrt{\text{MSE}}$ op een leeftijd van 96 maanden, een jaar voor de afknotleeftijd, nog steeds ongeveer 5000 kg FPCM. Dit geeft aan dat de modelvoorspellingen niet enkel voor referentiemoment B inaccuraat zijn, maar dat dit over het gehele beschouwde tijdsinterval het geval is.

EVI

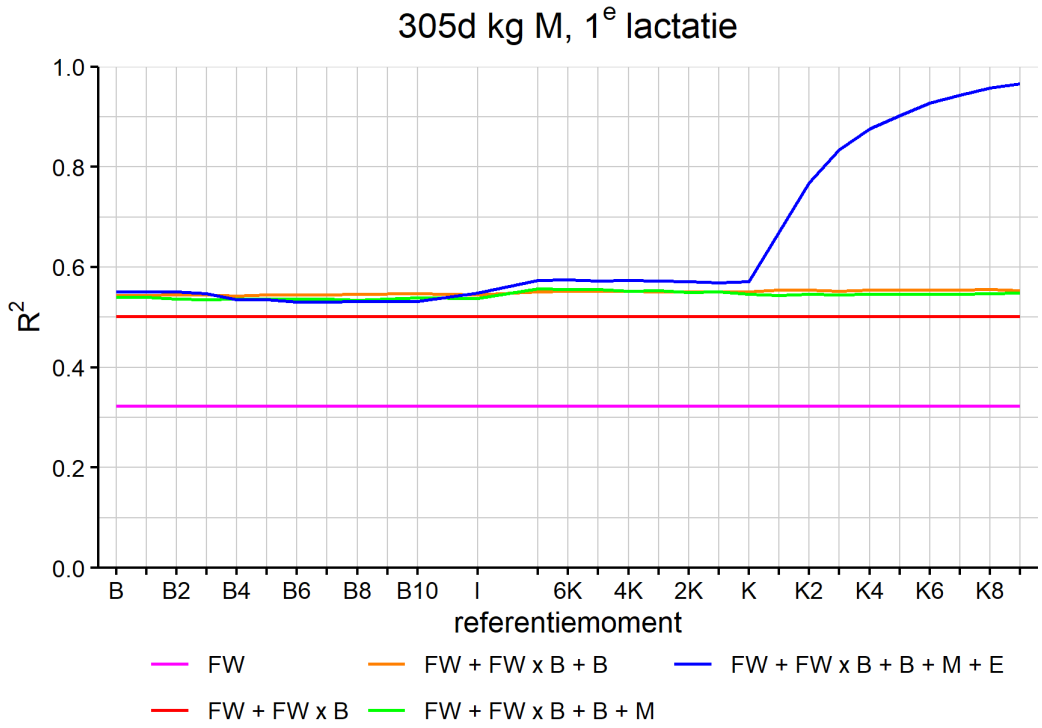
Wat betreft de EVI-modellen, kunnen gelijkaardige patronen waargenomen worden als voor de EVA-modellen, op twee kleine verschillen na. Een eerste verschil is waar te nemen op momenten vlak voordat het productiekengetal effectief wordt gerealiseerd: geen enkele machine learning techniek slaagt er in op deze momenten waarden voor $R^2_{landelijk}$ te realiseren hoger dan 0.95, dit als gevolg van de beperkte hoeveelheid waarnemingen die per bedrijfsspecifiek model beschikbaar is. Een tweede verschil is dat de performantiedip die voor tkt en de kg FPCM-productie op 3/4/5 jaar leeftijd kon worden waargenomen bij EVA, niet zichtbaar is bij de EVI-modellen. Dit is op zijn beurt te verklaren aan de hand van het feit dat de dieren die deze dip veroorzaken vaak op een beperkt aantal bedrijven gelokaliseerd zijn, waardoor de EVI-modellen of geen last hebben van deze dieren, of verschillende dergelijke dieren in de trainingsdataset hebben en daardoor geen onrealistische voorspellingen voor deze dieren maken.

3.4.4 Bijdrage van de verschillende informatiebronnen

In de inputdatasets kunnen vijf informatiebronnen onderscheiden worden: fokwaarden (FW), fokwaarde × bedrijf interactie-effect (FW × B), bedrijfsparameters (B), maternale eigenprestaties (M) en eigenprestaties (E). Om de bijdrage van de verschillende informatiebronnen te achterhalen, werden de EVA-modellen voor de 305d-productie tijdens de 1^e lactatie, uitgedrukt in kg M, een aantal keer opnieuw getraind en geëvalueerd, waarbij telkens een of meerdere van deze vijf informatiebronnen uit de dataset werden weggelaten. In Figuur 3.1 worden de resultaten van deze analyse

voor SVR_radial weergegeven, de resultaten voor de andere beschouwde machine learning technieken zijn nagenoeg identiek en worden daarom niet weergegeven.

Op Figuur 3.1 is te zien dat bij de geboorte hoofdzakelijk de fokwaarden en het fokwaarde × bedrijf interactie-effect verantwoordelijk zijn voor de gerealiseerde performantie. De stijging van $R^2_{landelijk}$ die kan waargenomen worden naarmate de tijd vordert, is daarentegen volledig te wijten aan het beschikbaar komen van informatie met betrekking tot de eigenprestaties.



Figuur 3.1: Bijdrage van de verschillende informatiebronnen tot de modelperformantie

Het basisniveau dat wordt behaald indien SVR_radial wordt toegepast op basis van een dataset met enkel en alleen de fokwaarden, ligt in lijn met wat theoretisch verwacht kan worden: vermenigvuldigt men de heritabiliteit van de 305d kg M-productie tijdens de eerste lactatie (0.48; CRV, 2020a) met de gemiddelde betrouwbaarheid van de fokwaarde kgM (79.8%), dan bekomt men 0.38, wat niet ver van $R^2_{landelijk}$ ligt die in Figuur 3.1 wordt gerapporteerd (0.32).

Opvallend is de grote bijdrage van het fokwaarde × bedrijf interactie-effect, welke een stijging van $R^2_{landelijk}$ teweegbrengt van 0.18. Worden daarnaast nog de bedrijfsparameters en de maternale eigenprestaties toegevoegd aan de dataset, dan leidt dit slechts tot een marginale stijging van $R^2_{landelijk}$. Met betrekking tot de maternale eigenprestaties betekent dit echter niet dat zij niet belangrijk zijn, zo heeft het beschikbaar komen van maternale eigenprestaties een invloed op de betrouwbaarheid

en de nominale waarde van de fokwaarden die haar dochter zal toegekend krijgen. Dit effect kan echter niet gedetecteerd worden op basis van de door CRV aangeleverde dataset, aangezien deze een momentopname is van (een deel van) de CRV-databank. zoals reeds aangehaald in §2.2, impliceert dit dat alle informatie die beschikbaar was op het moment dat de aangeleverde dataset werd samengesteld (dus ook de maternale eigenprestaties) werd gebruikt voor de fokwaardeberekeningen. Aangezien de informatie die kan gehaald worden uit de maternale eigenprestaties dus reeds verwerkt is in de fokwaarden, is het logisch dat het toevoegen van maternale eigenprestaties aan de inputdataset niet leidt tot een significante toename van $R^2_{landelijk}$.

3.4.5 Performantie op bedrijfsniveau

Finaal is het niet de performantie op landelijk niveau die telt, maar de performantie van de modellen op bedrijfsniveau: een model mag nog perfect inter-bedrijfsvariantie kunnen modelleren, als het niet in staat is om de intra-bedrijfsvariantie op een degelijke manier in kaart te brengen, zal de doorsnee gebruiker/melkveehouder het model als waardeloos beschouwen. Daarom wordt in Tabel 3.4 voor ieder beschouwd kengetal en voor iedere beschouwde strategie, de gemiddelde R^2-waarde op bedrijfsniveau ($\overline{R^2}_{bedrijf}$) weergegeven voor de best presterende machine learning technieken op referentiemoment B (geboorte). Aangezien $R^2_{bedrijf}$ sterk kan variëren tussen de verschillende bedrijven (zie figuren 3.2 - 3.4) mogen de waarden in Tabel 3.4 enkel maar gebruikt worden om trends te identificeren en niet om een absoluut getal te kleven op de performantie van een bepaald machine learning model.

Tabel 3.4 vertoont, niet onverwacht, eenzelfde patroon als Tabel 3.3, waarbij SVR_radial domineert bij EVA, lasso regressie en random forest de meest geschikte technieken zijn voor EVI en relatief eenvoudige machine learning technieken zoals linreg, ridge en PCA + lasso de beste performanties realiseren bij STIN. Ook de patronen die kunnen gezien worden als de R^2-waarden van de verschillende productiekengetallen onderling worden vergeleken zijn vergelijkbaar met deze in Tabel 3.3: van alle productiekengetallen met betrekking tot de eerste lactatie, zijn %V en %E het best voorspelbaar en wat betreft de kg FPCM-productie op x jaar leeftijd, vertoont $\overline{R^2}_{bedrijf}$ bij STIN een duidelijk stijgende trend.

Er is echter een opvallend verschil tussen Tabel 3.4 en Tabel 3.3: de waarden in Tabel 3.4 zijn zonder uitzondering kleiner dan de waarden die worden gerapporteerd in Tabel 3.3 Dit is niet geheel onverwacht en kan eenvoudig theoretisch onderbouwd worden op basis van de betekenis van R^2. Stellen we de verklaarde inter-bedrijfsvariantie gelijk aan $Var(A)$, de verklaarde intra-bedrijfsvariantie gelijk aan $Var(B)$ en de onverklaarde variantie op landelijk niveau gelijk aan $Var(\epsilon_{landelijk})$, waarbij we voor de

Tabel 3.4: $\overline{R^2}_{bedrijf}$ voor de beste machine learning modellen per strategie, geëvalueerd op referentiemoment B (geboorte). Machine learning technieken die een $\overline{R^2}_{bedrijf}$ realiseerden die minder dan 0.01 verschilde van deze van het beste model, worden weergegeven met een superscript

	EVA		**EVI**		**STIN**	
kengetal	**techniek**	$\overline{\mathbf{R^2}}$	**techniek**	$\overline{\mathbf{R^2}}$	**techniek**	$\overline{\mathbf{R^2}}$
1e lactatie						
305d kg M	SVR_radial	0.35	random forest[c]	0.23	PCA + lasso[ab]	0.41
305d kg V	SVR_radial	0.29	lasso[h]	0.18	PCA + lasso[ab]	0.36
305d kg E	SVR_radial	0.25	random forest	0.20	PCA + lasso[ab]	0.34
305d kg V+E	SVR_radial	0.25	random forest	0.18	PCA + lasso[ab]	0.33
305d %V	SVR_radial[i]	0.61	lasso	0.51	PCA + lasso[ab]	0.67
305d %E	SVR_radial	0.55	lasso	0.30	linreg[bg]	0.60
305d kg FCM	SVR_radial	0.26	random forest	0.17	PCA + lasso[ab]	0.34
305d kg FPCM	SVR_radial	0.26	random forest	0.18	PCA + lasso[ab]	0.34
tkt	ridge[acdg]	0.00	ridge	-0.03	PCA + lasso[abc]	0.06
x jaar lft.						
kg FPCM; 3 jaar	SVR_radial	0.16	ridge[h]	0.11	linreg[bg]	0.25
kg FPCM; 4 jaar	SVR_radial	0.19	random forest	0.13	linreg[bg]	0.29
kg FPCM; 5 jaar	SVR_radial[a]	0.16	random forest	0.11	linreg[bg]	0.32
kg FPCM; 6 jaar	SVR_radial	0.13	random forest	0.09	linreg[bg]	0.36
kg FPCM; 7 jaar	lasso[b]	0.14	random forest	0.07	PCA + lasso[b]	0.41
kg FPCM; 8 jaar	SVR_poly[c]	0.01	boosting tree[g]	0.07	lasso[g]	0.24
leven						
kg FPCM	SVR_radial[b]	0.01	random forest	0.01	PCA + lasso	0.29

[a] linreg; [b] ridge; [c] lasso; [d] PCA + linreg; [e] PCA_10; [f] PCA_100; [g] PCA + lasso; [h] random forest; [i] boosting tree; [k] SVR_radial; [m] SVR_poly

eenvoud veronderstellen dat A, B en $\epsilon_{landelijk}$ onafhankelijk zijn van elkaar, dan wordt R^2 (indien positief) op landelijk niveau bij benadering gegeven door:

$$R^2_{landelijk} = \frac{Var(A) + Var(B)}{Var(A) + Var(B) + Var(\epsilon_{landelijk})} \tag{3.1}$$

Wordt R^2 echter op bedrijfsniveau berekend, dan valt $Var(A)$ weg uit zowel de noemer als de teller en wordt $Var(\epsilon_{landelijk})$ vervangen door $Var(\epsilon_{bedrijf})$, waardoor R^2 (indien positief) op bedrijfsniveau bij benadering gegeven wordt door:

$$R^2_{bedrijf} = \frac{Var(B)}{Var(B) + Var(\epsilon_{bedrijf})} \tag{3.2}$$

Aangezien varianties per definitie positief zijn en het verschil tussen $Var(\epsilon_{landelijk})$ en $Var(\epsilon_{bedrijf})$ meestal relatief beperkt is, volgt uit vergelijkingen (3.1) en (3.2) dat, voor een gegeven model, steeds geldt dat $R^2_{landelijk} \geq R^2_{bedrijf}$.

Het verschil tussen de waarden gerapporteerd in Tabel 3.4 en Tabel 3.3 is hierbij opvallend groot, zo zijn de waarden voor $\overline{R^2}_{bedrijf}$ die in Tabel 3.4 worden gerapporteerd voor kg FPCM op 8 jaar leeftijd en de levensproductie, net niet gelijk aan nul, terwijl de corresponderende waarden voor $R^2_{landelijk}$ in Tabel 3.3 respectievelijk 0.59 en

0.30 bedroegen. Dit wijst er op dat de EVA- en EVI-modellen voor deze parameters vooral inter-bedrijfsvariantie verklaarden. Wat betreft de vergelijking tussen EVA en EVI, blijkt uit Tabel 3.4, analoog aan wat in Tabel 3.3 kon worden opgemerkt, dat ook voor $\overline{R^2}_{bedrijf}$ de mogelijkheid van de EVI-modellen om te focussen op de intra-bedrijfsvariantie niet doorweegt ten opzichte van het beperktere aantal records dat beschikbaar is om de modellen op te stellen.

Ook voor STIN zijn de R^2-waarden in Tabel 3.4 kleiner dan in Tabel 3.3, maar de verschillen zijn minder groot, waardoor in Tabel 3.4 de meerwaarde van STIN nog meer tot uiting komt dan in Tabel 3.3 Zo is het in Tabel 3.4 niet uitzonderlijk dat de gerapporteerde $\overline{R^2}_{bedrijf}$ voor STIN groter is dan de som van de $\overline{R^2}_{bedrijf}$-waarden voor EVA en EVI. Bij alle productiekengetallen vertoont STIN een grote meerwaarde tegenover EVA en EVI, maar vooral voor de levensproductie is de meerwaarde van STIN, met een stijging van $\overline{R^2}_{bedrijf}$ met 0.28, verrassend hoog.

Tabel 3.3 toont dus opnieuw aan dat de meerwaarde van machine learning technieken bij het voorspellen van productiekengetallen vooral gelegen is in het feit dat de verschillende technieken het probleem op een verschillende manier kunnen benaderen, waardoor na uitbreiding van de originele inputdataset met de modelpredicties van de individuele machine learning modellen (stacking) een sterke stijging van de modelperformantie kan gerealiseerd worden.

Invloed van het aantal waarnemingen op de modelperformantie

Zoals reeds eerder aangehaald, moeten de waarden voor $\overline{R^2}_{bedrijf}$ die in Tabel 3.4 worden gerapporteerd, voorzichtig geïnterpreteerd worden aangezien deze zeer sterk variëren tussen de verschillende bedrijven in de dataset. Om te kunnen bestuderen wat de invloed is van het aantal waarnemingen dat voor een bepaald bedrijf beschikbaar is op $R^2_{bedrijf}$, werd voor alle machine learning modellen die werden opgesteld bij het modelleren van de 305d kg FPCM-productie tijdens de eerste lactatie op referentiemoment B, een regressie-analyse uitgevoerd. Bij deze regressie-analyse werden de gerealiseerde waarden voor $R^2_{bedrijf}$ gemodelleerd in functie van het aantal waarnemingen dat per bedrijf beschikbaar was, met behulp van regressiemodel (3.3), met y_i de $R^2_{bedrijf}$ voor bedrijf i, x_i het aantal beschikbare waarnemingen voor bedrijf i en ϵ_i de foutterm.

$$y_i = a + b \log_{10}(x_i) + \epsilon_i \tag{3.3}$$

De coëfficiënten b die bij deze regressie-analyse werden bekomen, worden weergegeven in Tabel 3.5. Voor lasso regressie worden de uitgevoerde regressie-analyses daarnaast gevisualiseerd in figuren 3.2 - 3.4.

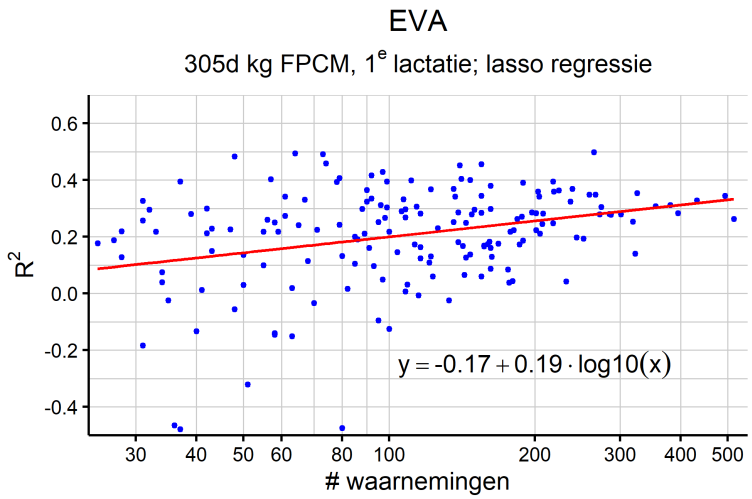
Tabel 3.5: Regressiecoëfficiënten b die werden bekomen bij het modelleren van $R^2_{bedrijf}$ in functie van het aantal waarnemingen per bedrijf met regressiemodel (3.3)

techniek	**EVA**	**EVI**	**STIN**
linreg	0.18		0.10
ridge	0.18	0.30	0.10
lasso	0.19	0.46	0.12
PCA + linreg	0.21		0.15
PCA_10		0.65	
PCA_100	0.26		0.14
PCA + ridge	0.18	0.30	0.10
PCA + lasso	0.18	0.37	0.10
random forest	0.16	0.23	0.13
boosting tree	0.14	0.31	0.16
SVR_radial	0.13	0.26	0.14
SVR_poly	0.22	0.29	0.14

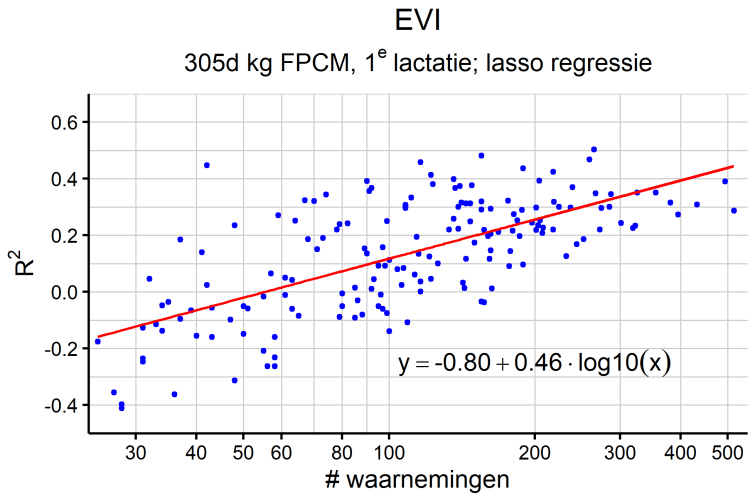
Wordt gekeken naar de resultaten in Tabel 3.5, dan kan gezien worden dat alle bekomen regressiecoëfficiënten b positief zijn, wat betekent dat voor alle modeltechnieken en voor alle strategieën een hoger aantal waarnemingen per bedrijf gepaard gaat met een hogere $R^2_{bedrijf}$. Tussen de verschillende machine learning technieken zijn er weinig opvallende verschillen, maar de bekomen regressiecoëfficiënten verschillen wel sterk tussen de verschillende strategieën. Hierbij is het zo dat de regressiecoëfficiënten b het hoogst zijn voor EVI, het laagst voor STIN en dat EVA zich ergens tussenin bevindt. Dit betekent dat bij de EVI-modellen de modelperformantie op bedrijfsniveau het gevoeligst is voor het aantal waarnemingen per bedrijf en dat bij de STIN-modellen de modelperformantie op bedrijfsniveau het minst gevoelig is voor variatie in het aantal waarnemingen per bedrijf. Dit kan ook gezien worden in figuren 3.2 - 3.4, waar duidelijk te zien is dat de regressierechte in Figuur 3.3 veel steiler verloopt dan deze in Figuur 3.4.

Voor EVI is het eigenlijk logisch dat bij deze techniek de helling van de regressierechte het grootst is, bedrijven met een beperkt aantal records kunnen bij EVI immers niet profiteren van de grotere hoeveelheid waarnemingen die op landelijk niveau beschikbaar is. Zeker voor kleine bedrijven met minder dan 50 waarnemingen leidt dit tot zeer lage R^2-waarden, die in de meeste gevallen zelfs negatief zijn.

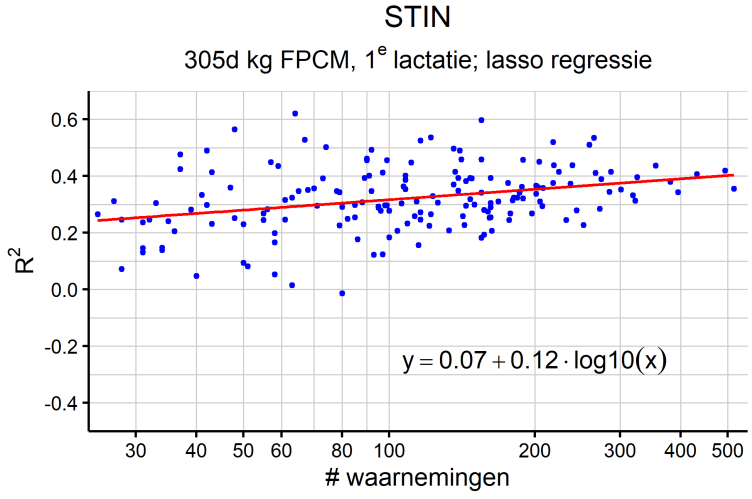
Aangezien bij EVA de bedrijven met minder waarnemingen kunnen profiteren van het feit dat er op landelijk niveau een grotere hoeveelheid waarnemingen beschikbaar is, is de helling van de regressierechte in Figuur 3.2 een stuk lager dan deze die bij EVI werd bekomen. Desondanks is $R^2_{bedrijf}$ op verschillende bedrijven met een beperkt aantal waarnemingen, nog steeds negatief.



Figuur 3.2: Invloed van het aantal waarnemingen per bedrijf op $R^2_{bedrijf}$ bij lasso regressie volgens EVA



Figuur 3.3: Invloed van het aantal waarnemingen per bedrijf op $R^2_{bedrijf}$ bij lasso regressie volgens EVI

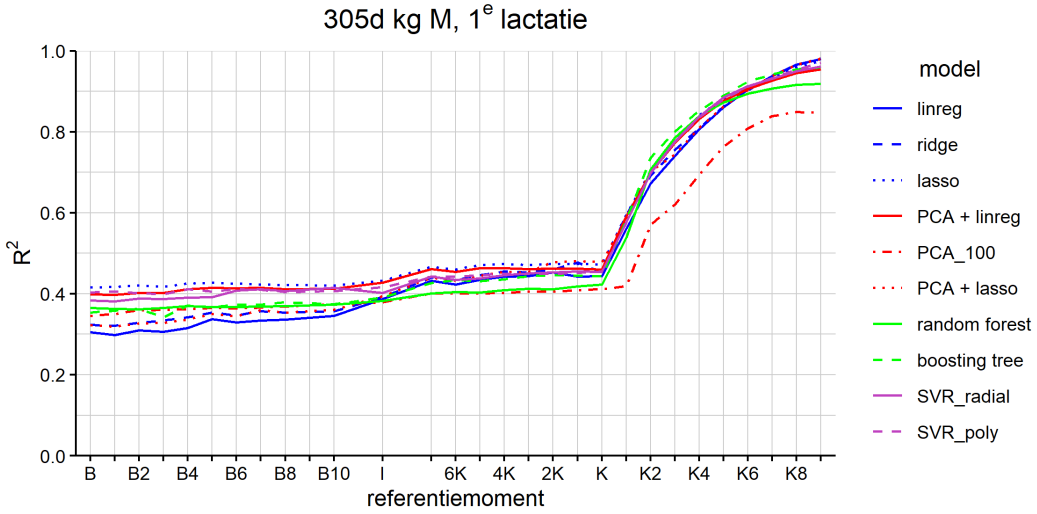


Figuur 3.4: Invloed van het aantal waarnemingen per bedrijf op $R^2_{bedrijf}$ bij lasso regressie volgens STIN

Ook hier biedt STIN meerdere voordelen, zo kan in Figuur 3.4 niet alleen gezien worden dat de gemiddelde waarde voor $R^2_{bedrijf}$ groter is dan voor EVA en EVI, maar ook dat er geen bedrijven meer zijn waar $R^2_{bedrijf}$ beduidend lager is dan nul. Als derde voordeel van STIN kan tot slot ook opgemerkt worden dat bij STIN voor een gegeven aantal waarnemingen per bedrijf, de variantie van $R^2_{bedrijf}$ kleiner is dan dat deze is voor EVA en EVI, waardoor op voorhand $R^2_{bedrijf}$ voor een bepaald bedrijf nauwkeuriger kan ingeschat worden.

3.4.6 Performantie op vooraf ongeziene bedrijven

In §2.3.9 werd vermeld dat bij het opstellen van het fokwaarde × bedrijf interactie-effect, in het achterhoofd gehouden werd dat de opgestelde (EVA-)modellen ook moesten kunnen toegepast worden op vooraf ongeziene bedrijven. Voor alle resultaten die hiervoor werden gerapporteerd en besproken, werd echter geen gebruik gemaakt van deze mogelijkheid. Om te evalueren in hoeverre de EVA-modellen in staat zijn om ook voor vooraf ongeziene bedrijven de productiekengetallen van melkkoeien te voorspellen, werd de hele procedure om de EVA machine learning modellen op te stellen en te evalueren herhaald voor de 305d kg M-productie tijdens de 1^e lactatie, maar de *folds* werden opgesteld op basis van het bedrijfsnummer en niet op basis van de levensnummers van de dieren, zodat bedrijven of in de testdataset, of in de trainingsdataset aanwezig waren, maar nooit in beide tegelijk. De waarden voor $R^2_{landelijk}$ die hierbij werden bekomen voor de verschillende machine learning technieken, worden weergegeven in Figuur 3.5



Figuur 3.5: R^2-waarden op landelijk niveau voor bedrijven die niet in de trainingsdataset werden opgenomen (EVA; 305d kg M, 1^e lactatie)

Er is te zien dat de machine learning modellen effectief in staat zijn om voorspellingen te maken voor bedrijven die niet werden opgenomen in de trainingsdataset,

maar dat $R^2_{landelijk}$ voor deze bedrijven lager is dan voor de bedrijven die wel werden opgenomen in de trainingsdataset (Figuur B.1). Desondanks realiseren de meeste EVA-modellen op refentiemoment B (geboorte) nog steeds hogere performanties dan deze die behaald kunnen worden indien men enkel en alleen de fokwaarden zou opnemen in de inputdataset (Figuur 3.1). Dit geeft aan dat de machine learning modellen in staat zijn om ook voor bedrijven die niet werden opgenomen in de trainingsdataset, het fokwaarde × bedrijf interactie-effect in rekening te brengen. Bovendien zijn de curves in Figuur 3.5 nog steeds relatief glad (*smooth*) wat er op wijst dat de mate waarin de EVA-modellen last hebben van overfitting als gevolg van 'identificatie' van bedrijven in de trainingsdataset, relatief beperkt is.

De opgestelde EVA-modellen kunnen dus gebruikt worden voor vooraf ongeziene bedrijven, maar aangezien de modelperformantie voor deze bedrijven lager is dan voor de bedrijven die wel werden opgenomen in de trainingsdataset, is het aangeraden om alle bedrijven waarvoor men voorspellingen wil maken, op te nemen in de trainingsdataset, net zoals dat voor de fokwaardeberekeningen wordt gedaan.

3.4.7 Invloed van melkregime, epigenetica en leeftijd bij eerste kalving

Als laatste aspect met betrekking tot de bespreking van de opgestelde machine learning methoden, werd bestudeerd wat de gevoeligheid van de opgestelde machine learning modellen is voor de parameters met betrekking tot het melkregime, de pariteit van de moeder en de leeftijd bij eerste kalving (LEK). Hiervoor werden de coëfficiënten a die werden bekomen bij linreg, ridge en lasso (strategie EVA) op passende referentiemomenten, via formule (3.4), omgerekend naar de gevoeligheden van de machine learning modellen voor de betreffende parameter.

$$\frac{\Delta(y-\overline{y})}{\sigma_y} = a\frac{\Delta(x-\overline{x})}{\sigma_x} \Leftrightarrow \frac{\Delta y}{\Delta x} = a\frac{\sigma_y}{\sigma_x} \tag{3.4}$$

Voor de variabelen met betrekking tot het melkregime en de pariteit van de moeder, werden de modelcoëfficiënten na omrekening volgens formule (3.4) nog verminderd met de waarden die werden bekomen voor de referentiesituatie (tweemaal daags melken en moeder was in pariteit nul tijdens de dracht) zodat de gevoeligheden die in Tabel 3.6 worden vermeld, steeds relatief worden uitgedrukt tegenover de referentiesituatie.

In Tabel 3.6 kan gezien worden dat driemaal daags melken leidt tot een 305d-productie tijdens de eerste lactatie die gemiddeld 750-950 kg M hoger ligt dan bij tweemaal daags melken. Dit komt, relatief uitgedrukt tegenover het landelijk gemiddelde, over-

een met een stijging van 9-11.5% De percentages vet (-0.2%) en de percentages eiwit (-0.1%) dalen echter iets, waardoor de stijging in geproduceerde hoeveelheid vet en eiwit iets kleiner is: + 15-23 kg V (+ 4-7 %) en + 17-25 kg E(+ 6-9 %). Deze resultaten liggen volledig in lijn met de de waarden die tijdens de literatuurstudie werden gevonden (§1.3.4). Koeien melken met een robot (code 9) heeft gelijkaardige effecten op de melkproductie, maar de grootte van de productiestijging is beduidend kleiner dan voor driemaal daags melken.

Wat betreft het effect van de pariteit van de moeder op de productiekengetallen van de eerste lactatie, lijken de gevoeligheden die in Tabel 3.6 gerapporteerd worden, de resultaten in Tabel 1.3 tegen te spreken: een hogere pariteit van de moeder lijkt gepaard te gaan met een hogere 305d-productie tijdens de eerste lactatie. De gevoeligheden die worden gerapporteerd in Tabel 3.6 zijn echter niet consistent, zo is het onwaarschijnlijk dat indien de moeder in de tweede pariteit was tijdens de dracht, dat het positief epigenetisch effect op de 305d-productie dubbel zo groot is dan indien de moeder in de derde pariteit of hoger was tijdens de dracht. Hierdoor kan op basis van Tabel 3.6 enkel maar besloten worden dat, als er een epigenetisch effect is van de pariteit van de moeder tijdens de dracht, dat dit effect relatief bescheiden is en zeker niet te vergelijken is met het effect van driemaal daags melken op de melkproductie.

Ook met betrekking tot het effect van de LEK, lijken de resultaten die in Tabel 3.6 gerapporteerd worden, de wetenschappelijke literatuur tegen te spreken (§1.3.3). Zo gaat een stijging van de LEK met een maand (30.5 dagen) volgens de opgestelde ridge regressie gepaard met een daling van de 305d-productie tijdens de eerste lactatie van 88,5 kg M ($-1.1 \cdot 30.5 - 1.8 \cdot 30.5$) terwijl in de wetenschappelijke literatuur steeds wordt vermeld dat een hogere LEK gepaard gaat met een hogere 305d-productie tijdens de eerste lactatie.

De waarneming dat de gevoeligheden van de machine learning modellen niet altijd overeenkomen met wat in de wetenschappelijke literatuur wordt vermeld, heeft twee mogelijke oorzaken. Een eerste mogelijke oorzaak is dat er in de dataset die aangeleverd werd door CRV, effectief indicaties zijn dat een hogere pariteit van de moeder tijdens de dracht en een lagere LEK gepaard gaan met een hogere melkproductie tijdens de eerste lactatie. Een tweede, meer plausibele, mogelijke oorzaak is dat de opgestelde machine learning modellen niet geschikt zijn om het effect van deze factoren te bestuderen. Het hoofddoel van de ontwikkelde machine learning modellen was dan ook om productiekengetallen te voorspellen en niet om het effect van de LEK of de pariteit van de moeder tijdens de dracht te bestuderen. Indien men het effect van invloedsfactoren met een (vermoedelijk) beperkt effect op de melkproductie wenst te achterhalen, is het dus aangeraden om (machine learning) modellen te gebruiken die speciaal voor dit doel ontworpen werden.

Tabel 3.6: Gevoeligheid van enkele machine learning modellen (EVA) voor parameters in de inputdataset m.b.t. het melkregime, de pariteit van de moeder tijdens de dracht en de LEK. De gevoeligheden m.b.t het melkregime en de pariteit van de moeder werden bepaald op referentiemoment B, de gevoeligheden m.b.t de LEK, werden bepaald op referentiemoment 7K

kengetal	techniek	melkregime			pariteit moeder				vruchtbaarheid		
		2	3	9	0	1	2	3+	dgn 1	ins n	dgn n
1e lactatie											
305d kg M	linreg	0	956	243	0	19	55	26	-1.1	-6.5	-1.7
	ridge	0	935	253	0	19	55	26	-1.1	-6.5	-1.8
	lasso	0	748	232	0	17	52	21	-1.0	-5.8	-1.7
305d kg V	linreg	0.0	22.8	7.4	0.0	3.4	5.3	3.7	-0.05	-0.26	-0.04
	ridge	0.0	21.6	7.9	0.0	3.4	5.3	3.7	-0.05	-0.25	-0.04
	lasso	0.0	15.0	7.8	0.0	3.4	5.3	3.6	-0.04	-0.24	-0.04
305d kg E	linreg	0.0	25.2	5.5	0.0	-0.3	1.61	0.4	-0.05	-0.17	-0.04
	ridge	0.0	24.1	5.9	0.0	-0.3	1.6	0.4	-0.05	-0.18	-0.04
	lasso	0.0	17.2	4.9	0.0	-0.4	1.5	0.2	-0.05	-0.17	-0.04
305d kg V+E	linreg	0.0	47.8	12.8	0.0	3.2	6.9	4.2	-0.10	-0.43	-0.08
	ridge	0.0	45.6	13.8	0.0	3.1	6.9	4.2	-0.10	-0.43	-0.08
	lasso	0.0	32.0	12.7	0.0	3	6.7	3.8	-0.09	-0.40	-0.08
305d %V	linreg	0.00	-0.19	-0.03	0.00	0.03	0.03	0.03	0.00	0.00	0.00
	ridge	0.00	-0.2	-0.03	0.00	0.03	0.03	0.03	0.00	0.00	0.00
	lasso	0.00	-0.19	-0.02	0.00	0.03	0.03	0.03	0.00	0.00	0.00
305d %E	linreg	0.00	-0.09	-0.03	0.00	-0.02	-0.01	-0.01	0.00	0.00	0.00
	ridge	0.00	-0.1	-0.03	0.00	-0.02	-0.01	-0.01	0.00	0.00	0.00
	lasso	0.00	-0.1	-0.03	0.00	-0.02	-0.01	-0.01	0.00	0.00	0.00
305d kg FCM	linreg	0	725	209	0	59	101	66	-1.11	-6.50	-1.28
	ridge	0	697	220	0	59	101	66	-1.10	-6.40	-1.29
	lasso	0	527	210	0	57	100	62	-1.04	-5.89	-1.21
305d kg FPCM	linreg	0	737	200	0	44	89	53	-1.18	-6.18	-1.29
	ridge	0	710	212	0	44	90	54	-1.18	-6.19	-1.31
	lasso	0	532	198	0	43	87	50	-1.11	-5.77	-1.24
tkt	linreg	0	15.3	8.9	0	2.9	5.9	1.6	-0.02	0.76	0.03
	ridge	0	10.4	6.7	0	3.1	5.8	1.9	0.00	1.06	0.02
	lasso	0	11.9	7.0	0	2.9	5.7	1.8	0.00	1.15	0.00

3.5 Praktijktoepassing: afvoer van overtollig jongvee

Als sluitstuk van deze masterproef werd bestudeerd in hoeverre de opgestelde machine learning modellen een meerwaarde zouden kunnen bieden bij het afvoeren van overtollig vrouwelijk jongvee. De bijdrage die de opgestelde machine learning technieken in deze context zouden kunnen leveren, werd hierbij geanalyseerd aan de hand van drie criteria: (1) percentage van de afgevoerde dieren dat correct wordt afgevoerd (2) percentage van de 50% beste dieren dat wordt behouden en (3) de selectierespons.

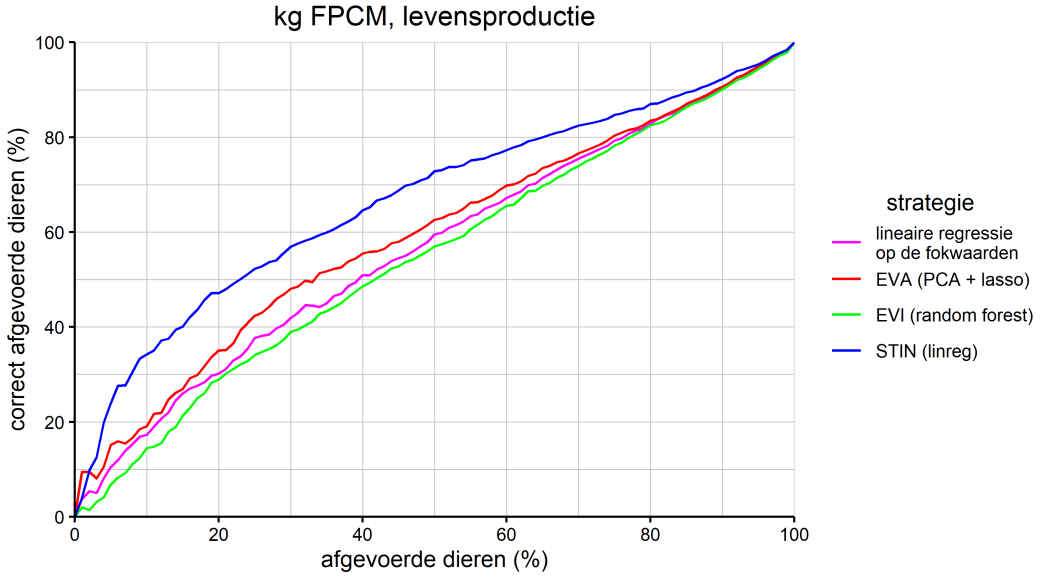
Voor het bepalen van de nominale waarden van deze criteria voor een gegeven productiekengetal en afvoerpercentage (x%) werden hierbij eerst voor alle dieren enerzijds de modelvoorspellingen voor referentiemoment B en anderzijds de gerealiseerde waarden voor het beschouwde productiekengetal verzameld. Vervolgens werden op bedrijfsniveau telkens de x% slechtste dieren geselecteerd, dit zowel voor de modelvoorspellingen als voor de werkelijke productiekengetallen. Er werd hierbij gekozen om op bedrijfsniveau te werken, omdat werken op landelijk niveau zou inhouden dat van bepaalde bedrijven al het jongvee zou worden afgevoerd, wat niet realistisch is. Als per bedrijf de dieren waren geselecteerd die in aanmerking kwamen om af te voeren, werd op landelijk niveau berekend (1) welk percentage van de afgevoerde dieren correct werd afgevoerd (2) welk percentage van de 50% beste dieren aangehouden werd (enkel voor afvoerpercentages $\leq$ 50%) en (3) hoeveel het landelijk gemiddelde voor het beschouwde kengetal steeg (of daalde in het geval van de tkt). De respons van het landelijk gemiddelde op het afvoeren van x% van het jongvee, stemt hierbij overeen met de respons die gemiddeld verwacht kan worden op bedrijfsniveau.

De resultaten van deze analyse worden voor verschillende relevante kengetallen weergegeven in Bijlage C. De selectierespons wordt hierbij zowel absoluut als relatief weergegeven, waarbij voor het berekenen van de relatieve selectierespons werd gedeeld door de maximaal haalbare selectierespons. De maximaal haalbare selectierespons stemt hierbij overeen met de selectierespons die zou bekomen worden bij een perfect afvoerbeleid, waarbij dieren zouden worden afgevoerd op basis van de werkelijke kengetallen die ze zouden realiseren en niet op basis van voorspellingen.

In Bijlage C worden de resultaten steeds weergegeven voor selectie op referentiemoment B, wat praktisch zou overeenstemmen met het selecteren van de af te voeren vaarskalveren van zodra de genomic fokwaarden beschikbaar zijn. In de praktijk wordt vaak iets langer gewacht aangezien men zeker wil zijn dat er (1) dat jaar zeker voldoende vaarskalveren geboren zullen worden en (2) dat er op korte termijn niet

een nog slechter vaarskalf geboren wordt. Aangezien de performantie van de meeste modellen enkel maar toeneemt naarmate de dieren ouder worden, kan echter verwacht worden dat ook de eventuele meerwaarde van machine learning technieken enkel maar zal toenemen naarmate de dieren ouder worden, zodat de nominale waarden van de evaluatiecriteria op referentiemoment B als een baseline geïnterpreteerd kunnen worden.

Aangezien de meeste melkveehouders vooral de vaarskalveren willen afvoeren die een lage levensproductie zullen realiseren, wordt in wat volgt voornamelijk gefocust op dit productiekengetal. Aangezien bovendien binnen eenzelfde strategie de ranking van de verschillende machine learning technieken, zoals enigszins verwacht kon worden, vrij analoog is aan de ranking van de verschillende technieken voor $\overline{R^2}_{bedrijf}$, wordt ook hier niet dieper op ingegaan.

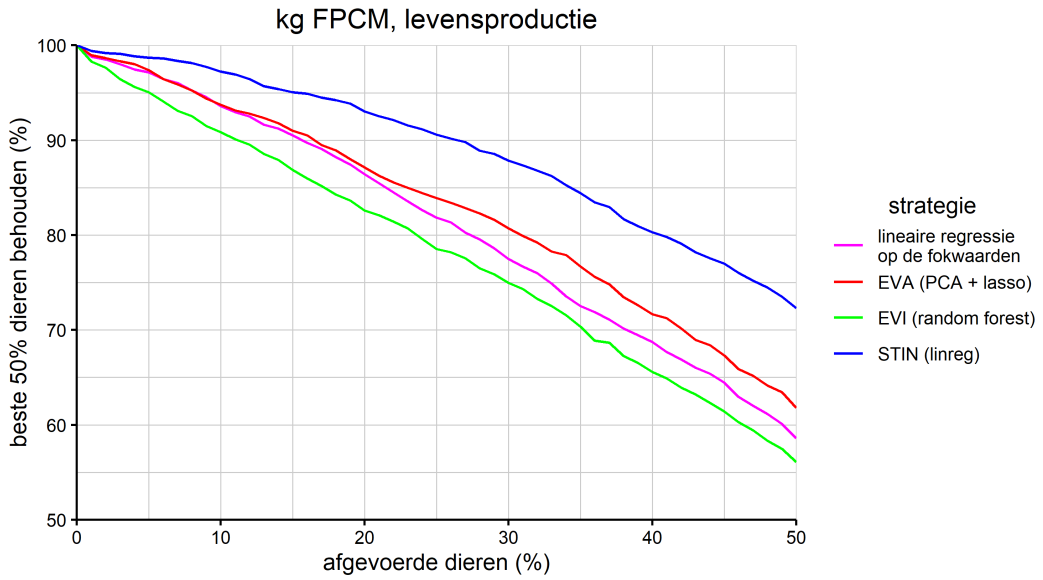


Figuur 3.6: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM

In Figuur 3.6 wordt het percentage correct afgevoerde dieren vergeleken voor de verschillende beschouwde strategieën (EVA, EVI en STIN). Als referentie wordt daarnaast ook de curve voor multiple lineaire regressie op enkel en alleen de fokwaarden op landelijk niveau weergegeven. Om Figuur 3.6 goed te kunnen interpreteren is het belangrijk om te weten hoe de curves voor een willekeurig en een perfect afvoerbeleid verlopen. De curve voor een willekeurig afvoerbeleid start in de oorsprong en loopt vervolgens via de eerste bissectrice tot in de rechterbovenhoek van Figuur 3.6 De curve voor een perfect afvoerbeleid start ook in de oorsprong en eindigt ook in de rechterbovenhoek van Figuur 3.6, maar neemt tussen dit begin- en eindpunt steeds een waarde van 100% aan. Slechte selectiemethodes leveren dus een curve op die dicht bij de eerste bissectrice ligt, terwijl de curves voor goede selectiemethodes

meer naar de linkerbovenhoek van Figuur 3.6 bewegen. De curve voor het percentage correct afgevoerde dieren vertoont aldus zeer sterke gelijkenissen met ROC-curves, die vaak in de medische wetenschappen worden gebruikt om de sensitiviteit en de specificiteit van testen te evalueren. Analoog aan hoe dit voor ROC-curves gebeurt, kan de kwaliteit van een bepaalde selectiemethode dan ook geëvalueerd worden aan de hand van de oppervlakte onder de curve (AUC, *Area Under the Curve*). Dit werd voor alle opgestelde machine learning modellen gedaan, waarna, om de figuren overzichtelijk te houden, enkel de beste machine learning techniek per strategie werd weerhouden om te visualiseren in figuren 3.6 - 3.9

In Figuur 3.6 kan gezien worden dat EVI veruit de minst performante strategie is, die nog slechter presteert dan een eenvoudige lineaire regressie op de fokwaarden, welke op zich ook niet echt een performante selectiemethodologie blijkt te zijn. EVA levert iets betere resultaten op dan lineaire regressie op de fokwaarden, maar de meerwaarde is niet echt spectaculair. De lage meerwaarde van de EVA- en EVI-modellen tegenover een willekeurig afvoerbeleid hoeft niet te verbazen, gegeven $\overline{R^2}_{bedrijf}$ die beide strategieën noteerden in Tabel 3.4 De situatie ligt echter helemaal anders voor STIN, die met eenvoudige lineaire regressie (voortbouwend op meer complexe EVA- en EVI-modellen) overduidelijk een grote meerwaarde biedt in het kader van het afvoeren van overtollig jongvee.

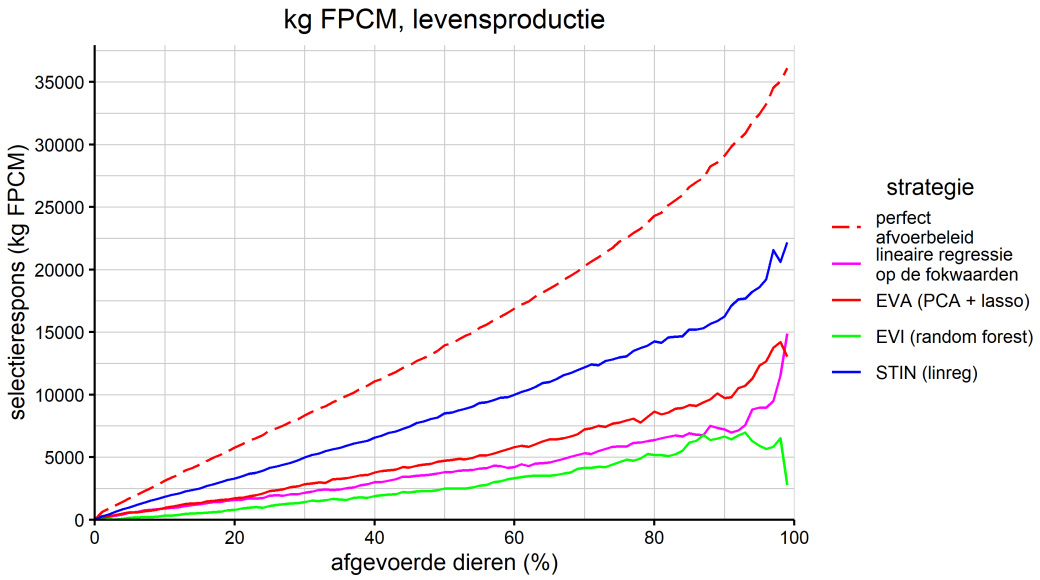


Figuur 3.7: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM

In Figuur 3.7 wordt het percentage van de beste 50% dieren dat wordt behouden na selectie van de af te voeren dieren, weergegeven. In deze grafiek resulteert een willekeurig afvoerbeleid in een rechte die van de linkerbovenhoek naar de rechterbe-

nedenhoek loopt, terwijl een perfect afvoerbeleid een horizontale lijn oplevert die de y-as snijdt op 100%. Ook hier kan dus de kwaliteit van de selectiemethodologie geëvalueerd worden aan de hand van de oppervlakte onder de curve. Analoog aan wat uit Figuur 3.6 kon besloten worden, kan ook in Figuur 3.7 worden gezien dat (1) EVI een slechtere selectiemethodologie oplevert dan lineaire regressie op de fokwaarden, (2) de meerwaarde van EVA tegenover lineaire regressie op de fokwaarden beperkt is en (3) STIN duidelijk een meerwaarde biedt in het kader van het afvoeren van overtollig jongvee.

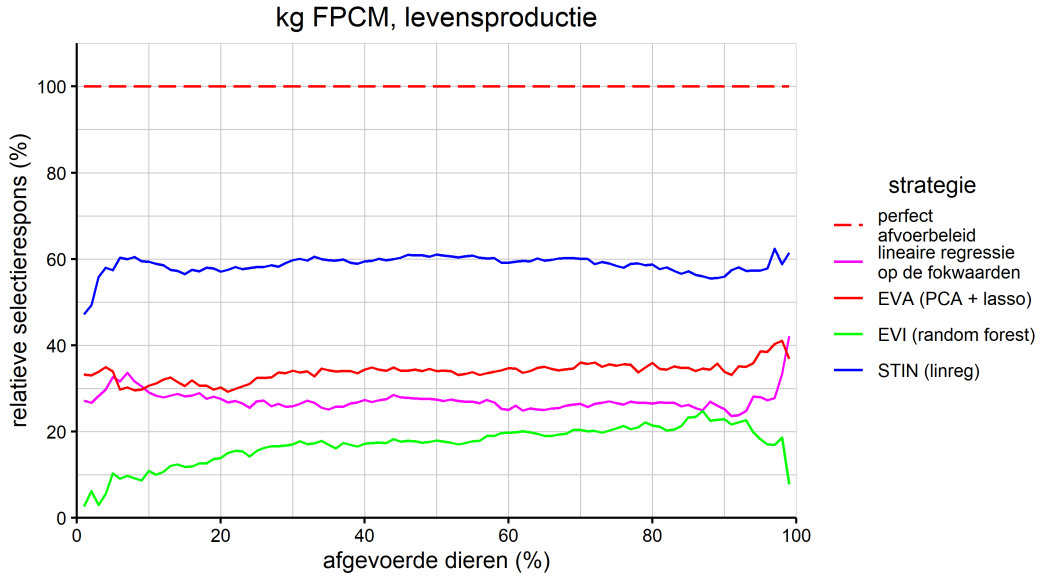
Uit de combinatie van Figuur 3.6 en Figuur 3.7 kan veel interessante informatie gehaald worden over welke dieren nu effectief zullen worden afgevoerd bij een gegeven afvoerpercentage. Wordt bijvoorbeeld 20% van het jongvee afgevoerd, dan kan voor STIN in Figuur 3.6 afgelezen worden dat 47% van de afgevoerde dieren correct wordt afgevoerd en in Figuur 3.7 kan afgelezen worden dat 93% van de beste 50% dieren wordt behouden. Wordt deze informatie gecombineerd, dan kan eenvoudig worden afgeleid dat de 20% afgevoerde dieren zullen samengesteld zijn uit 9.4% dieren die tot de slechtste 20% van de dieren behoren, 3.5% dieren die tot de beste 50% van de dieren behoren en 7.1% dieren die een hogere levensproductie zullen realiseren dan de 20% slechtste dieren, maar een lagere levensproductie dan de beste 50% van de dieren.



Figuur 3.8: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM

Uit figuren 3.6 en 3.7 blijkt dus dat STIN een meerwaarde biedt in het correct identificeren van af te voeren jongvee. Dit vertaalt zich in figuur 3.9 in het feit dat de relatieve selectierespons die op vlak van levensproductie kan gerealiseerd worden met STIN ongeveer 25 procentpunten hoger ligt dan deze voor EVA en ruim dubbel

zo groot is als deze die kan bereikt worden met behulp van lineaire regressie op de fokwaarden. Opvallend in Figuur 3.9 en alle figuren in Bijlage C is dat de curves voor de relatieve selectierespons steeds relatief vlak zijn.



Figuur 3.9: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM

Indien de waarden voor $\overline{R^2}_{bedrijf}$ die worden gerapporteerd in Tabel 3.4 worden vergeleken met de waarden voor de relatieve selectierespons die worden gevisualiseerd in Bijlage C, dan kan gezien worden dat $\overline{R^2}_{bedrijf}$ niet spectaculair hoog moet zijn om een goede relatieve selectierespons te realiseren. Zo levert een $\overline{R^2}_{bedrijf}$ van 0.29 voor STIN reeds een relatieve selectierespons op van ongeveer 60% (Figuur C.60). Dit is eigenlijk niet onlogisch, aangezien het volstaat om de goede van de slechte (of minder goede) dieren te kunnen onderscheiden om reeds een aanvaardbare relatieve selectierespons te bekomen.

Wordt in Figuur 3.8 gekeken naar de nominale waarde van de selectierespons, dan kan gezien worden dat de selectierespons die met STIN kan behaald worden, niet verwaarloosbaar klein is: indien een melkveehouder 20% van zijn vrouwelijk jongvee afvoert op basis van de STIN-voorspellingen voor levensproductie, dan kan verwacht worden dat de gemiddelde levensproductie met 3300 kg FPCM zal stijgen, wat overeenkomt met een stijging van 7.4% tegenover het landelijk gemiddelde. De selectierespons die kan verwacht worden, loopt overigens snel op naarmate hogere afvoerpercentages worden gehanteerd, indien bijvoorbeeld 30% van het aanwezige jongvee wordt afgevoerd op basis van de STIN-voorspellingen, dan is de selectierespons reeds 5000 kg FPCM, wat ongeveer overeenkomt met een halve lactatieproductie.

Op basis van bovenstaande waarnemingen zou men (foutief) kunnen afleiden dat, indien men het afvoerpercentage opdrijft door minder gebruikskruisingen toe te passen, men de gemiddelde levensproductie op bedrijfsniveau makkelijk met meerdere procenten zou kunnen verhogen. Indien bijvoorbeeld het percentage af te voeren vrouwelijk jongvee zou stijgen van 15% naar 30%, dan stijgt de selectierespons die met STIN behaald kan worden voor de levensproductie, van 2500 kg FPCM naar 5000 kg FPCM. Een foutieve redenering hierbij zou zijn dat men dus kan veronderstellen dat de gemiddelde levensproductie van het aangehouden vrouwelijk jongvee door dit verhoogd afvoerpercentage met 2500 kg FPCM zal toenemen. Er zit echter een addertje onder het gras: de dieren die bij de referentiesituatie geïnsemineerd zouden worden met een stier van een vleesras, hebben doorgaans een lager genetisch niveau dan de dieren die met een Holstein stier worden geïnsemineerd. De kalveren die uit deze koeien worden geboren, zullen als gevolg hiervan gemiddeld ook een lager genetisch niveau hebben. Aangezien de fokwaarden en het fokwaarde × bedrijf interactie-effect een belangrijke rol spelen bij het voorspellen van productiekengetallen (zie §3.4.4) kan dan ook verwacht worden dat door minder gebruikskruisingen toe te passen, de gemiddelde levensproductie van de geboren vaarskalveren (indien deze allemaal zouden worden aangehouden) lager zal zijn dan in de referentiesituatie waar meer gebruikskruisingen worden gebruikt. Het basisniveau van waar men vertrekt, verlaagt dus indien men het afvoerpercentage zou opdrijven door meer gebruikskruisingen in te zetten. Hierdoor zal het opdrijven van het afvoerpercentage van 15% naar 30% hoogstwaarschijnlijk wel leiden tot een iets hoger levensproductie van de dieren die worden aangehouden, maar de additionele respons zal lager zijn dan de 2500 kg FPCM die men intuïtief uit Figuur 3.8 zou kunnen afleiden. Hoeveel de additionele selectierespons dan wel is die kan gehaald worden door het aandeel gebruikskruisingen te verlagen, werd niet onderzocht in deze masterproef, maar vormt een interessant uitgangspunt voor eventueel vervolgonderzoek, waarbij gekeken kan worden naar hoe machine learning technieken kunnen gebruikt worden om het fokkerijbeleid op bedrijfsniveau te optimaliseren.

HOOFDSTUK 4

BESLUIT

Op basis van de resultaten van deze masterproef, kan besloten worden dat bij een voldoende hoog aantal waarnemingen, de meerwaarde van individuele machine learning technieken om productiekengetallen bij melkvee te voorspellen, relatief beperkt is. Machine learning methoden die in staat zijn om interacties tussen de verschillende variabelen in de dataset in rekening te brengen, zoals support vector regressie met een radiale kernel, leveren meestal iets hogere performanties op dan multiple lineaire regressie, maar de meerwaarde is steeds relatief beperkt. Worden machine learning methoden echter als aanvulling op lineaire regressie gebruikt en via stacking gecombineerd tot één predictief model, dan bieden machine learning methoden wél een aanzienlijke meerwaarde bij het voorspellen van productiekengetallen bij melkvee. Bijvoorbeeld bij het productiekengetal levensproductie, uitgedrukt in kg FPCM, kon via stacking een R^2-waarde van 0.47 gerealiseerd worden, een stijging van maar liefst 0.17 tegenover de R^2-waarde die het best presterende individuele machine learning model realiseerde.

Bij een beperkte hoeveelheid waarnemingen, indien men bijvoorbeeld voor een individueel bedrijf een predictief model wil opstellen, zijn lasso regressie en random forest regressie de meest geschikte machine learning methoden om productiekengetallen bij melkvee te voorspellen. Doordat deze technieken over de mogelijkheid beschikken om predictieve modellen op te stellen op basis van datasets met meer variabelen dan waarnemingen, bieden deze machine learning technieken een groot voordeel tegenover multiple lineaire regressie. Bij multiple lineaire regressie kan immers vaak maar een deel van alle beschikbare informatie gebruikt worden doordat deze techniek relatief gevoelig is voor overfitting. De modelperformantie die kan gerealiseerd worden met een bedrijfsspecifiek machine learning model is als gevolg van de beperkte hoeveelheid waarnemingen echter steeds lager dan deze die kan behaald worden met landelijke modellen, zelfs al gebruikt men machine learning technieken zoals random forest regressie of lasso regressie.

Machine learning modellen staan echter niet op zichzelf: ieder goed model begint met een gepaste primaire dataverwerking. Ook in deze masterproef bleek herhaaldelijk dat de primaire dataverwerking een groot aandeel had in de gerealiseerde model-

performantie, zo resulteerde bijvoorbeeld het op een doordachte wijze inbrengen van een fokwaarde × bedrijf interactie-effect en het aggregeren van alle mpr-data tot een vast aantal variabelen, in hogere modelperformanties.

Ondanks het feit dat een passende primaire dataverwerking en stacking van machine learning technieken voor de meeste productiekengetallen resulteerde in R^2-waarden van ±0.50 of hoger, bleven de predictiefouten echter meestal erg groot. De relatief grote predictiefouten staan echter geen praktijktoepassingen van de opgestelde machine learning modellen in de weg, zo kon aangetoond worden dat machine learning modellen een grote meerwaarde kunnen bieden bij het afvoeren van overtollig jongvee met het oog op een verhoogde levensproductie. In vergelijking met de selectierespons die hierbij kan worden behaald met multiple lineaire regressie op de fokwaarden, slaagde het beste machine learning model er zelfs in om deze selectierespons te verdubbelen.

BIBLIOGRAFIE

Aguilar, I., Misztal, I., Johnson, D. L., Legarra, A., Tsuruta, S., & Lawlor, T. J. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *Journal of Dairy Dcience*, 93:743–752.

Bannister, A. J. & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, 21:381–395.

Bernier-Dodier, P., Delbecchi, L., Wagner, G. F., Talbot, B. G., & Lacasse, P. (2010). Effect of milking frequency on lactation persistency and mammary gland remodeling in mid-lactation cows. *Journal of Dairy Science*, 93:555–564.

Berry, D. P. & Cromie, A. R. (2009). Associations between age at first calving and subsequent performance in Irish spring calving Holstein–Friesian dairy cows. *Livestock Science*, 123:44–54.

Bouraoui, R., Lahmar, M., Majdoub, A., & Belyea, R. (2002). The relationship of temperature-humidity index with milk production of dairy cows in a mediterranean climate. *Animal Research*, 51:479–491.

Braunschweig, M., Jagannathan, V., Gutzwiller, A., & Bee, G. (2012). Investigations on transgenerational epigenetic response down the male line in f2 pigs. *PloS one*, 7(2):e30583.

Bryant, J. R., López-Villalobos, N., Pryce, J. E., Holmes, C. W., & Johnson, D. L. (2007). Quantifying the effect of thermal environment on production traits in three breeds of dairy cattle in New Zealand. *New Zealand Journal of Agricultural Research*, 50(3):327–338.

Campos, M. S., Wilcox, C. J., Head, H. H., Webb, D. W., & Hayen, J. (1994). Effects on production of milking three times daily on first lactation Holsteins and Jerseys in Florida. *Journal of Dairy Science*, 77(3):770–773.

Champagne, F., Diorio, J., Sharma, S., & Meaney, M. J. (2001). Naturally occurring variations in maternal behavior in the rat are associated with differences in estrogen-inducible central oxytocin receptors. *Proceedings of the National Academy of Sciences*, 98(22):12736–12741.

Champagne, F. A., Weaver, I. C. G., Diorio, J., Dymov, S., Szyf, M., & Meaney, M. J. (2006). Maternal care associated with methylation of the estrogen receptor-α1b promoter and estrogen receptor-α expression in the medial preoptic area of female offspring. *Endocrinology*, 147(6):2909–2915.

Christensen, O. F. & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genetics Selection Evolution*, 42:2.

Clark, D. A., Phyn, C. V. C., Tong, M. J., Collis, S. J., & Dalley, D. E. (2006). A systems comparison of once-versus twice-daily milking of pastured dairy cows. *Journal of Dairy Science*, 89:1854–1862.

Cockett, N. E., Jackson, S. P., Shay, T. L., Farnir, F., Berghmans, S., Snowder, G. D., Nielsen, D. M., & Georges, M. (1996). Polar overdominance at the ovine callipyge locus. *Science*, 273:236–238.

Colombani, C., Croiseau, P., Fritz, S., Guillaume, F., Legarra, A., Ducrocq, V., & Robert-Granié, C. (2012). A comparison of partial least squares (PLS) and sparse PLS regressions in genomic selection in French dairy cattle. *Journal of Dairy Science*, 95:2120–2131.

CRV (1998). E2; Lactatieproductie en 305 dagenproductie.

CRV (2015). E3; Netto Opbrengst en Lactatiewaarde.

CRV (2020a). E7; Fokwaardeschatting melkproductiekenmerken met testdagmodel.

CRV (2020b). Jaarstatistieken 2019 voor Nederland.

CVB (2016). Tabellenboek veevoeding 2016; voedernormen rundvee, schapen, geiten en voederwaarden voedermiddelen voor herkauwers.

Davis, M. S., Mader, T. L., Holt, S. M., & Parkhurst, A. M. (2003). Strategies to reduce feedlot cattle heat stress: Effects on tympanic temperature. *Journal of Animal Science*, 81(3):649–661.

Dikmen, S. & Hansen, P. J. (2009). Is the temperature-humidity index the best indicator of heat stress in lactating dairy cows in a subtropical environment? *Journal of Dairy Science*, 92:109–116.

Druet, T., Ahariz, N., Cambisano, N., Tamma, N., Michaux, C., Coppieters, W., Charlier, C., & Georges, M. (2014). Selection in action: dissecting the molecular underpinnings of the increasing muscle mass of Belgian Blue Cattle. *BMC genomics*, 15:796.

Dunn, R. J. H., Mead, N. E., Willett, K. M., & Parker, D. E. (2014). Analysis of heat stress in UK dairy cattle and impact on milk yields. *Environmental Research Letters*, 9:064006.

Feil, R. (2006). Environmental and nutritional effects on the epigenetic regulation of genes. *Mutation Research*, 600:46–57.

Francis, D., Diorio, J., Liu, D., & Meaney, M. J. (1999). Nongenomic transmission across generations of maternal behavior and stress responses in the rat. *Science*, 286:1155–1158.

Friggens, N. C., Emmans, G. C., & Veerkamp, R. F. (1999). On the use of simple ratios between lactation curve coefficients to describe parity effects on milk production. *Livestock Production Science*, 62:1–13.

Geeks for Geeks (2019). Stacking in machine learning. url: `https://www.geeksforgeeks.org/stacking-in-machine-learning/`. (geraadpleegd 18 december 2019).

González-Recio, O. (2012). Epigenetics: a new challenge in the post-genomic era of livestock. *Frontiers in genetics*, 2:106.

González-Recio, O., Ugarte, E., & Bach, A. (2012). Trans-generational effect of maternal lactation during pregnancy: a Holstein cow model. *PLoS One*, 7(12):e51816.

Grala, T. M., Phyn, C. V. C., Kay, J. K., Rius, A. G., Littlejohn, M. D., Snell, R. G., & Roche, J. R. (2011). Temporary alterations to milking frequency, immediately post-calving, modified the expression of genes regulating milk synthesis and apoptosis in the bovine mammary gland. In *Proceedings of the New Zealand Society of Animal Production*, volume 71, pages 3–8. New Zealand Society of Animal Production.

Guo, Z. & Swalve, H. H. (1995). Modelling of the lactation curve as a sub-model in the evaluation of test day records. In *Proceedings of the Interbull annual meeting*, volume 11. Interbull.

Hall, I. M., Shankaranarayana, G. D., Noma, K.-i., Ayoub, N., Cohen, A., & Grewal, S. I. S. (2002). Establishment and maintenance of a heterochromatin domain. *Science*, 297:2232–2237.

Hansen, J. V., Friggens, N. C., & Højsgaard, S. (2006). The influence of breed and parity on milk yield, and milk yield acceleration curves. *Livestock Science*, 104:53–62.

Hastie, T., Tibshirani, R., & Friedman, J. (2008). *The Elements of Statistical Learning*. Springer, second edition.

Henderson, C. R. (1973). Sire evaluation and genetic trends. *Journal of Animal Science*, 1973:10–41.

Henderson, C. R. (1984). *Applications of linear models in animal breeding*. University of Guelph.

Heslot, N., Yang, H.-P., Sorrells, M. E., & Jannink, J.-L. (2012). Genomic selection in plant breeding: a comparison of models. *Crop Science*, 52:146–160.

Hietanen, H. & Ojala, M. (1995). Factors affecting body weight and its association with milk production traits in finnish ayrshire and friesian cows. *Acta Agriculturae Scandinavica A-Animal Sciences*, 45:17–25.

Ho, L. & Crabtree, G. R. (2010). Chromatin remodelling during development. *Nature*, 463:474–484.

Holmes, C. W., Wilson, G. F., Mackenzie, D. D. S., & Purchas, J. (1992). The effects of milking once-daily throughout lactation on the performance of dairy cows grazing on pasture. In *Proceedings-New Zealand Society of Animal Production*, volume 52, pages 13–16. New Zealand Society of Animal Production.

Honarvar, M. & Ghiasi, H. (2013). A comparison of genomic predictions using support vector machines (SVMs) and GBLUP methods. *Agrochimica Research*, 57:3–21.

Jia, S., Noma, K.-i., & Grewal, S. I. S. (2004). RNAi-independent heterochromatin nucleation by the stress-activated ATF/CREB family proteins. *Science*, 304:1971–1976.

Jirtle, R. L. & Weidman, J. R. (2007). Imprinted and more equal. *American Scientist*, 95:143–149.

Keller, C. & Bühler, M. (2013). Chromatin-associated ncRNA activities. *Chromosome Research*, 21:627–641.

Keverne, E. B. & Curley, J. P. (2004). Vasopressin, oxytocin and social behaviour. *Current Opinion in Neurobiology*, 14:777–783.

Klei, L. R., Lynch, J. M., Barbano, D. M., Oltenacu, P. A., Lednor, A. J., & Bandler, D. K. (1997). Influence of milking three times a day on milk quality. *Journal of Dairy Science*, 80:427–436.

Lee, J.-Y. & Kim, I.-H. (2006). Advancing parity is associated with high milk production at the cost of body condition and increased periparturient disorders in dairy herds. *Journal of Veterinary Science*, 7(2):161–166.

Legarra, A., Aguilar, I., & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *Journal of Dairy Science*, 92:4656–4663.

Li, Z. & Sillanpää, M. J. (2012). Overview of lasso-related penalized regression methods for quantitative trait mapping and genomic selection. *Theoretical and Applied Genetics*, 125:419–435.

Lin, C. Y., McAllister, A. J., Batra T., R., Lee A., J., Roy G., L., Vesely J., A., Wauthy J., M., & Winter K., A. (1986). Production and reproduction of early and late bred dairy heifers. *Journal of Dairy Science*, 69:760–768.

Littlejohn, M. D., Walker, C. G., Ward, H. E., Lehnert, K. B., Snell, R. G., Verkerk, G. A., Spelman, R. J., Clark, D. A., & Davis, S. R. (2009). Effects of reduced frequency of milk removal on gene expression in the bovine mammary gland. *Physiological Genomics*, 41:21–32.

Long, N., Gianola, D., Rosa, G. J. M., & Weigel, K. A. (2011). Application of support vector regression to genome-assisted prediction of quantitative traits. *Theoretical and Applied Genetics*, 123:1065–1074.

Luan, T., Woolliams, J. A., Lien, S., Kent, M., Svendsen, M., & Meuwissen, T. H. E. (2009). The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. *Genetics*, 183(3):1119–1126.

Lumey, L. H. (1992). Decreased birthweights in infants after maternal in utero exposure to the Dutch famine of 1944–1945. *Paediatric and Perinatal Epidemiology*, 6(2):240–253.

Macdonald, K. A., Penno, J. W., Bryant, A. M., & Roche, J. R. (2005). Effect of feeding level pre-and post-puberty and body weight at first calving on growth, milk production, and fertility in grazing dairy cows. *Journal of Dairy Science*, 88:3363–3375.

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157:1819–1829.

Moser, G., Tier, B., Crump, R. E., Khatkar, M. S., & Raadsma, H. W. (2009). A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide snp markers. *Genetics Selection Evolution*, 41:56.

National Research Council (1971). *A guide to environmental research on animals*. National Academies.

Neugebauer, N., Räder, I., Schild, H. J., Zimmer, D., & Reinsch, N. (2010). Evidence for parent-of-origin effects on genetic variability of beef traits. *Journal of Animal Science*, 88:523–532.

Nor, N. M., Steeneveld, W., Van Werven, T., Mourits, M. C. M., & Hogeveen, H. (2013). First-calving age and first-lactation milk production on Dutch dairy farms. *Journal of Dairy Science*, 96:981–992.

Ogutu, J. O., Piepho, H.-P., & Schulz-Streeck, T. (2011). A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proceedings*, 5:S11.

Ogutu, J. O., Schulz-Streeck, T., & Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. *BMC proceedings*, 6:S10.

Okut, H., Wu, X.-L., Rosa, G. J. M., Bauck, S., Woodward, B. W., Schnabel, R. D., Taylor, J. F., & Gianola, D. (2013). Predicting expected progeny difference for marbling score in Angus cattle using artificial neural networks and Bayesian regression models. *Genetics Selection Evolution*, 45:34.

Oldenbroek, J. K. (1989). Parity effects on feed intake and feed efficiency in four dairy breeds fed ad libitum two different diets. *Livestock Production Science*, 21:115–129.

Piepho, H.-P. (2009). Ridge regression and extensions for genomewide selection in maize. *Crop Science*, 49:1165–1176.

Pintus, M. A., Gaspa, G., Nicolazzi, E. L., Vicario, D., Rossoni, A., Ajmone-Marsan, P., Nardone, A., Dimauro, C., & Macciotta, N. P. P. (2012). Prediction of genomic breeding values for dairy traits in Italian Brown and Simmental bulls using a principal component approach. *Journal of Dairy Science*, 95:3390–3400.

Pirlo, G., Miglior, F., & Speroni, M. (2000). Effect of age at first calving on production traits and on difference between milk yield returns and rearing costs in Italian Holsteins. *Journal of Dairy Science*, 83:603–608.

Pollott, G. E. (2000). A biological approach to lactation curve analysis for milk yield. *Journal of Dairy Science*, 83:2448–2458.

Ravagnolo, O. & Misztal, I. (2000). Genetic component of heat stress in dairy cattle, parameter estimation. *Journal of Dairy Science*, 83:2126–2130.

Rémond, B., Pomiès, D., Dupont, D., & Chilliard, Y. (2004). Once-a-day milking of multiparous Holstein cows throughout the entire lactation: milk yield and composition, and nutritional status. *Animal Research*, 53:201–212.

Scholtz, M. M., Van Zyl, J. P., & Theunissen, A. (2014). The effect of epigenetic changes on animal production. *Applied Animal Husbandry & Rural Development*, 7:7–10.

Schutz, M. M., Hansen, L. B., Steuernagel, G. R., & Kuck, A. L. (1990). Variation of milk, fat, protein, and somatic cells for dairy cattle. *Journal of Dairy Science*, 73:484–493.

Sinecen, M. (2019). Comparison of genomic best linear unbiased prediction and Bayesian regularization neural networks for genomic selection. *IEEE Access*, 7:79199–79210.

Singh, K., Molenaar, A. J., Swanson, K. M., Gudex, B., Arias, J. A., Erdman, R. A., & Stelwagen, K. (2012). Epigenetics: a possible role in acute and transgenerational regulation of dairy cow milk production. *Animal*, 6(3):375–381.

Smith, J. W., Ely, L. O., Graves, W. M., & Gilson, W. D. (2002). Effect of milking frequency on DHI performance measures. *Journal of Dairy Science*, 85:3526–3533.

Solberg, T. R., Sonesson, A. K., Woolliams, J. A., & Meuwissen, T. H. E. (2009). Reducing dimensionality for prediction of genome-wide breeding values. *Genetics Selection Evolution*, 41:29.

Sorensen, A., Muir, D. D., & Knight, C. H. (2001). Thrice-daily milking throughout lactation maintains epithelial integrity and thereby improves milk protein quality. *Journal of Dairy Research*, 68:15–25.

Stelwagen, K., Farr, V. C., Davis, S. R., & Prosser, C. G. (1995). Egta-induced disruption of epithelial cell tight junctions in the lactating caprine mammary gland. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 269:R848–R855.

Stelwagen, K. & Knight, C. H. (1997). Effect of unilateral once or twice daily milking of cows on milk yield and udder characteristics in early and late lactation. *Journal of Dairy Research*, 64:487–494.

Stelwagen, K., Phyn, C. V. C., Davis, S. R., Guinard-Flament, J., Pomies, D., Roche, J. R., & Kay, J. K. (2013). Invited review: reduced milking frequency: milk production and management implications. *Journal of Dairy Science*, 96:3401–3413.

Triantaphyllopoulos, K. A., Ikonomopoulos, I., & Bannister, A. J. (2016). Epigenetics and inheritance of phenotype variation in livestock. *Epigenetics & Chromatin*, 9:31.

Usai, M. G., Goddard, M. E., & Hayes, B. J. (2009). Lasso with cross-validation for genomic selection. *Genetics Research*, 91:427–436.

Van der Waaij, E. H., Galesloot, P. J. B., & Garrick, D. J. (1997). Some relationships between weights of growing heifers and their subsequent lactation performances. *New Zealand Journal of agricultural research*, 40(1):87–92.

Van Raden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, 91:4414–4423.

Volpe, T. A., Kidner, C., Hall, I. M., Teng, G., Grewal, S. I. S., & Martienssen, R. A. (2002). Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science*, 297:1833–1837.

Wall, E. H. & McFadden, T. B. (2007). Optimal timing and duration of unilateral frequent milking during early lactation of dairy cows. *Journal of Dairy Science*, 90:5042–5048.

West, J. W., Mullinix, B. G., & Bernard, J. K. (2003). Effects of hot, humid weather on milk temperature, dry matter intake, and milk yield of lactating dairy cows. *Journal of Dairy Science*, 86:232–242.

Wilmink, J. B. M. (1987). Adjustment of test-day milk, fat and protein yield for age, season and stage of lactation. *Livestock Production Science*, 16:335–348.

Wood, P. D. P. (1967). Algebraic model of the lactation curve in cattle. *Nature*, 216:164–165.

Yadav, M. P., Katpatal, B. G., & Kaushik, S. N. (1977). Components of inverse polynomial function of lactation curve, and factors affecting them in Hariana and its Friesian crosses. *Indian Journal of Animal Sciences*, 47(12):777–781.

Young, L. J., Wang, Z., Donaldson, R., & Rissman, E. F. (1998). Estrogen receptor α is essential for induction of oxytocin receptor by estrogen. *Neuroreport*, 9:933–936.

BIJLAGE A

REGRESSIEBOOM-GEBASEERDE MACHINE LEARNING METHODEN

A.1 Regressieboom

Bij het modelleren van y in functie van $\boldsymbol{x} = [x_1, x_2, ..., x_p, ...x_P]^T$ met behulp van een regressieboom, wordt de p-dimensionale predictor-ruimte (*predictor space*) gepartitioneerd in K regio's $R_1, R_2, ..., R_k, ..., R_K$, zodat iedere waarneming y_i kan worden gemodelleerd aan de hand van (A.1), met $\overline{y}_{R_k}$ het gemiddelde van alle y_i's waarvoor geldt dat $\boldsymbol{x_i} \in R_k$

$$\hat{y}_i = \begin{cases} \overline{y}_{R_1} & \text{als } \boldsymbol{x_i} \in R_1 \\ \overline{y}_{R_2} & \text{als } \boldsymbol{x_i} \in R_2 \\ \vdots & \\ \overline{y}_{R_k} & \text{als } \boldsymbol{x_i} \in R_k \\ \vdots & \\ \overline{y}_{R_K} & \text{als } \boldsymbol{x_i} \in R_K \end{cases} \tag{A.1}$$

Het opstellen van de partitie van de predictor-ruimte gebeurt hierbij op een sequentiële wijze en via eenvoudige lineaire criteria, zodat voor iedere $\boldsymbol{x_i}$ aan de hand van een beslissingsboom (= regressieboom) kan bepaald worden tot welke regio $\boldsymbol{x_i}$ behoort. De verliesfunctie die wordt gebruikt bij het opstellen van de partitie van de predictor-ruimte wordt gegeven door (A.2), welke equivalent is aan de *quadratic loss* functie (1.9) die wordt gebruikt bij multiple lineaire regressie.

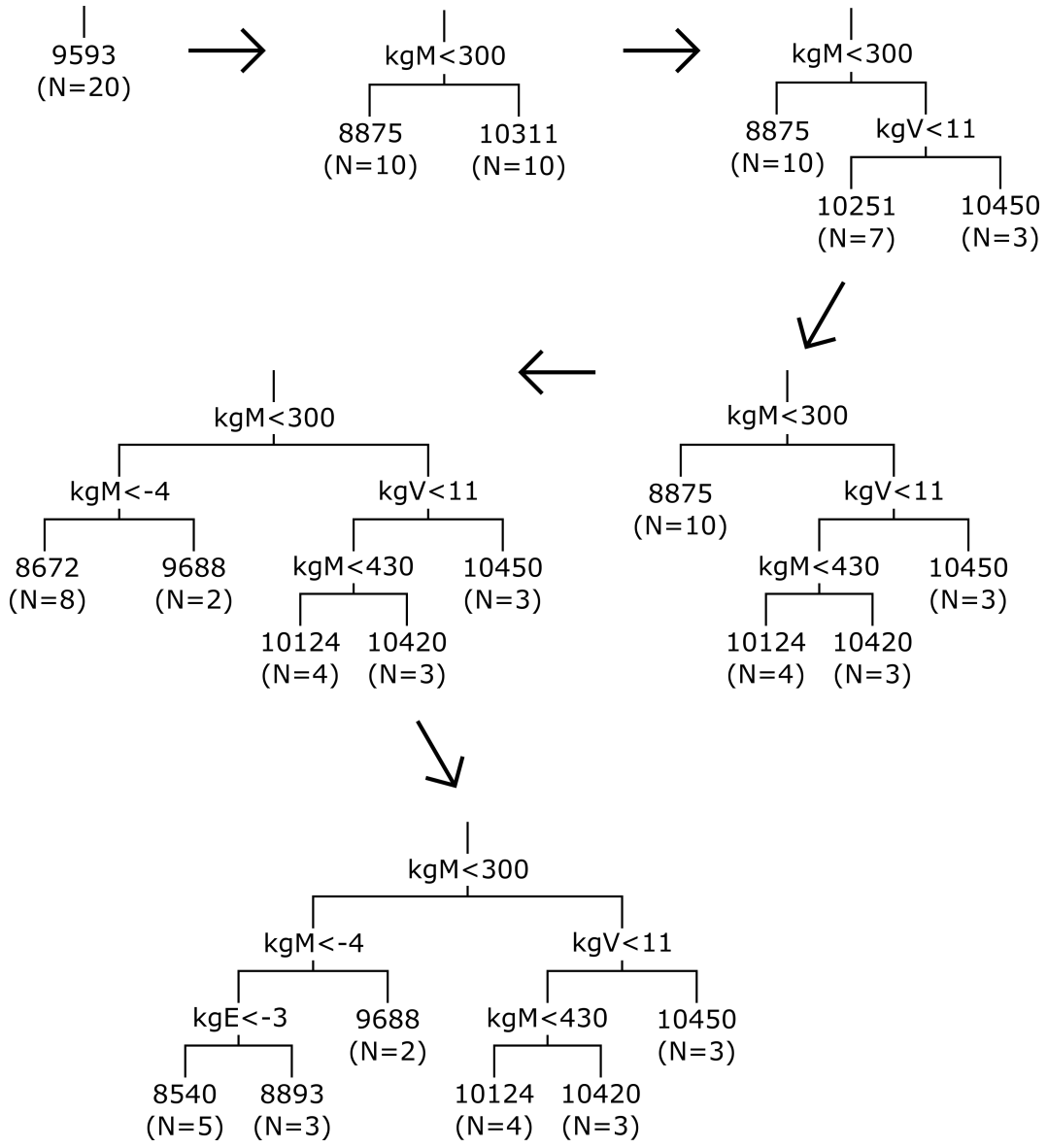
$$L = \sum_{k=1}^{K} \sum_{i \in R_k} (y_i - \overline{y}_{R_k})^2 \tag{A.2}$$

Bij het opstellen van een regressieboom wordt vertrokken van een triviale regressieboom waarbij de hele predictor-ruimte wordt ondergebracht in één regio R_1. Vervolgens wordt iteratief uit alle mogelijke combinaties van een k, p en s, die combinatie bepaald die leidt tot een maximale reductie van verliesfunctie (A.2) indien regio R_k in twee nieuwe regio's wordt opgesplitst volgens het lineair criterium $x_p < s$ (regio $R_{k'}$ indien waar, regio R_{K+1} indien onwaar) Om te vermijden dat deze procedure zou leiden tot een partitie van de predictor-ruimte waarbij iedere regio R_k maar één $\boldsymbol{x_i}$ zou omvatten, worden bij iedere iteratie enkel maar die k's in overweging genomen waarvoor geldt dat het aantal waarnemingen $\boldsymbol{x_i}$ in regio R_k groter is dan een vrij te kiezen thresholdwaarde m. De iteratieve procedure stopt hierdoor van zodra er geen enkele regio R_k meer is met meer dan m waarnemingen. In Figuur A.1 wordt deze iteratieve procedure geïllustreerd voor het modelleren van de 305d kg FPCM-productie tijdens de eerste lactatie op basis van de fokwaarden kgM, kgV en kgE (Tabel 2.1)

A.2 Random forest regressie

Regressiebomen hebben verschillende voordelen die deze modeltechniek aantrekkelijk maken om te gebruiken wanneer meer klassieke modeltechnieken zoals lineaire regressie slechte performanties realiseren. Een eerste voordeel is hun robuustheid: de voorspellingen zullen nooit lager zijn dan de minimale waarde in de dataset en nooit hoger dan de maximale waarde in de dataset. Een tweede voordeel is het feit dat zij, ondanks hun relatief eenvoudige opbouw, makkelijk interacties tussen de verschillende variabelen in rekening kunnen brengen. Een derde voordeel tot slot is dat regressiebomen weinig tot geen data-*preprocessing* vragen, zo kunnen regressiebomen bijvoorbeeld perfect met discrete variabelen werken zonder dat deze moeten omgezet worden in dummievariabelen. Regressiebomen hebben echter een groot nadeel: hun vaak grote predictiefout.

Bij random forest regressie wordt dit probleem aangepakt via *bagging*, een techniek waarbij verschillende (vaak meerdere honderden) regressiebomen opgesteld worden op basis van de dataset, waarna de finale modelvoorspelling wordt gegeven door het gemiddelde van de modelvoorspellingen van al deze individuele regressiebomen. Het idee achter *bagging* is vergelijkbaar aan hoe het verzamelen van meer waarnemingen leidt tot een nauwkeurigere schatting van het gemiddelde. In het kader van random forest regressie daalt de predictiefout echter zelden tot nul, maar stabiliseert deze zich meestal bij een toenemend aantal regressiebomen op een waarde lager dan de gemiddelde predictiefout voor de individuele regressiebomen. Steeds de volledige dataset gebruiken en te werk gaan zoals in §A.1 werd beschreven, zou echter telkens dezelfde regressieboom opleveren, wat weinig zoden aan de dijk zou brengen.



Figuur A.1: Illustratie van de procedure die gevolgd wordt bij het opstellen van een regressieboom ($m = 5$). Indien aan een criterium voldaan is, wordt bij de eerstvolgende splitsing naar links gegaan. Aan het einde van iedere tak wordt de modelvoorspelling $\overline{y}_{R_k}$ weergegeven, samen met het aantal waarnemingen (N) binnen regio R_k

Daarom worden in de procedure twee random stappen ingebouwd om te verzekeren dat de verschillende bekomen regressiebomen niet identiek zouden zijn.

Een eerste random stap in de procedure, is dat niet alle waarnemingen worden gebruikt voor het opstellen van een individuele regressieboom: om de trainingsdataset voor een individuele regressieboom op te stellen, wordt uit de n beschibare waarnemingen, met terugplaatsing, n keer een waarneming gesampled. Dit leidt er toe dat bepaalde waarnemingen uit de originele dataset niet zullen worden gebruikt, terwijl andere waarnemingen meer dan één keer zullen voorkomen in de trainingsdataset voor een individuele regressieboom.

Een tweede random stap in de procedure is dat bij iedere iteratie waarbij een regio wordt opgesplitst in twee nieuwe regio's, maar een fractie van de P beschikbare variabelen wordt beschouwd. Hoe groot deze fractie is, is een hyperparameter van het random forest algoritme. Vaak gekozen waarden voor deze fractie zijn $\sqrt{P}/P$ en $(log_2 P)/P$.

A.3 Boosted regressiebomen

Analoog aan random forest regressie, worden ook bij boosted regressiebomen meerdere individuele regressiebomen opgesteld. Bij boosted regressiebomen worden de verschillende individuele regressiebomen (aantal = B) echter niet parallel maar sequentieel opgesteld, waarbij iedere individuele regressieboom wordt opgesteld op basis van de 'residuen' van het model uit de vorige stap. Hierdoor zal bij het opstellen van regressieboom b voornamelijk aandacht besteed worden aan het verklaren van de variatie die door de $b-1$ eerder opgestelde regressiebomen niet verklaard kon worden.

Initieel wordt vertrokken van het triviale model $\hat{y}_i = 0$ en komen de 'residuen' r_i dus overeen met de waarnemingen die voor de onafhankelijke variabelen beschikbaar zijn: $r_i = y_i$. Vervolgens wordt een eerste individuele regressieboom ($b = 1$) opgesteld die deze 'residuen' tracht te modelleren in functie van de onafhankelijke variabelen. Bij het opstellen van deze regressieboom wordt maximaal d keer een regio opgesplitst, zodat de opgestelde regressieboom maximaal $d+1$ regio's R_k omvat. Na het opstellen van deze eerste regressieboom wordt het initiële triviale model geüpdatet tot (A.3) en worden de 'residuen' geüpdatet volgens (A.4), met $\hat{y}_i^{(1)}$ de modelvoorspellingen van de eerste opgestelde regressieboom.

$$\hat{y}_i = \lambda \hat{y}_i^{(1)} \tag{A.3}$$

$$r_i^{(1)} = y_i - \lambda \hat{y}_i^{(1)} \tag{A.4}$$

in de tweede stap wordt de eerste procedure herhaald, maar wordt een regressieboom opgesteld om $r_i^{(1)}$ te modelleren in functie van de onafhankelijke variabelen. Ook hier wordt het maximaal aantal regio's weer beperkt tot $d+1$. Na het opstellen van deze tweede regressieboom wordt model (A.3) geüpdated tot (A.5) en worden de residuen herrekend volgens (A.6), met $\hat{y}_i^{(2)}$ de modelvoorspellingen van de tweede opgestelde regressieboom.

$$\hat{y}_i = \lambda \left[\hat{y}_i^{(1)} + \hat{y}_i^{(2)} \right] \tag{A.5}$$

$$r_i^{(2)} = r_i^{(1)} - \lambda \hat{y}_i^{(2)} \tag{A.6}$$

Deze tweede stap wordt vervolgens nog $B-2$ keer herhaald, waarbij na iedere stap het model wordt geüpdated tot (A.7) en de residuen worden herrekend volgens (A.8), met $\hat{y}_i^{(b)}$ de modelvoorspellingen van de de b-de regressieboom.

$$\hat{y}_i = \lambda \left[\hat{y}_i^{(1)} + \hat{y}_i^{(2)} + \ldots + \hat{y}_i^{(b)} \right] \tag{A.7}$$

$$r_i^{(b)} = r_i^{(b-1)} - \lambda \hat{y}_i^{(b)} \tag{A.8}$$

Nadat op deze manier B individuele regressiebomen werden opgesteld, wordt het finale model gegeven door (A.9), met $\hat{y}_i^{(b)}$ de modelvoorspelling van de b-de regressieboom.

$$\hat{y}_i = \lambda \sum_{b=1}^{B} \hat{y}_i^{(b)} \tag{A.9}$$

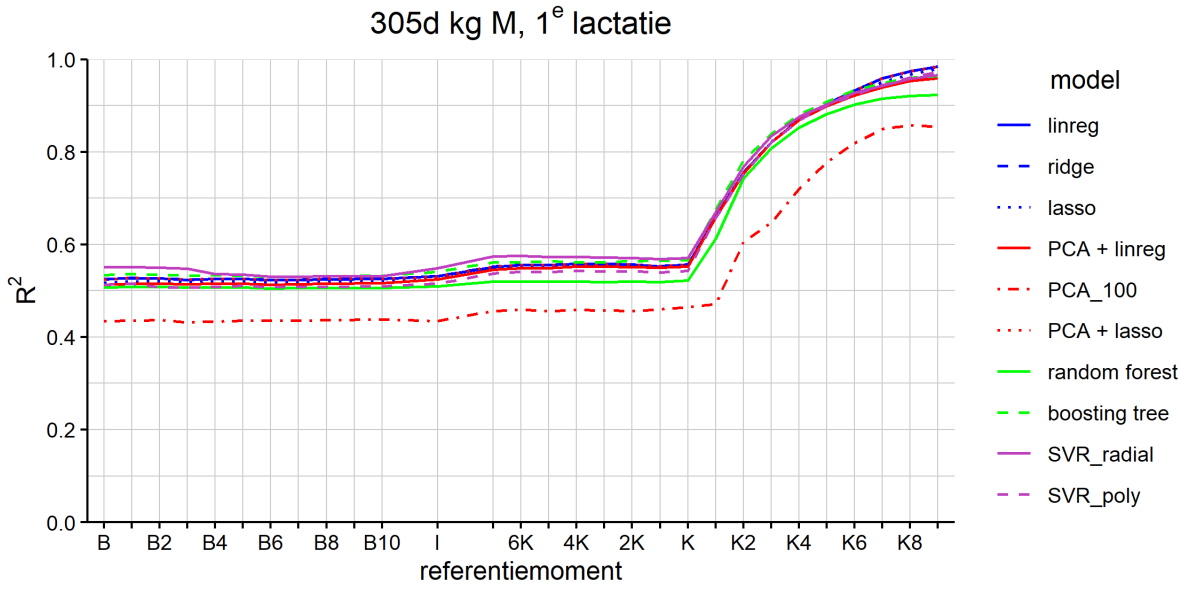
Zoals uit de beschrijving hiervoor afgeleid kan worden, hebben boosted regressiebomen 3 hyperparameters: B, d en λ. d legt vast hoe complex de individuele regressiebomen mogen zijn en bepaalt op die manier welk type interactie-effecten de boosted regressiebomen in rekening kunnen brengen. Indien $d = 1$ kunnen geen interactie-effecten tussen de verschillende variabelen in rekening gebracht worden, indien $d = 2$ kunnen interactie-effecten tussen twee variabelen in rekening gebracht worden,... Meestal is hierbij een relatief kleine waarde van d (1-3) reeds voldoende om goede modelperformanties te behalen. B is het aantal individuele regressiebomen dat wordt opgesteld. Bij random forest regressie heeft het aantal individuele regressiebomen dat wordt opgesteld niet echt veel invloed op de modelperformantie, zolang het aantal opgestelde individuele regressiebomen maar groot genoeg is. Doordat de individuele regressiebomen onafhankelijk van elkaar worden opgesteld, is de random forest-methodologie immers niet gevoelig aan overfitting, waardoor het weinig uitmaakt of er nu 500 of 50 000 individuele regressiebomen worden opgesteld. Boosted regressiebomen zijn echter wel gevoelig aan overfitting, als gevolg van het feit dat de individuele regressiebomen één na één worden opgesteld, waarbij in stap b de residuen $r_i^{(b-1)}$ worden gemodelleerd. Meestal is de mate van overfitting bij grote waarden van B echter relatief beperkt, zodat het meestal niet neerkomt op 10 regressiebomen meer of minder om de modelperformantie te maximaliseren. Tot slot is er nog λ, die ook wel de *learning rate* wordt genoemd en bepaalt hoe *greedy* het algoritme zich bij het opstellen van de boosted regressiebomen gedraagt. Een kleinere λ levert hierbij doorgaans een beter finaal model op, maar heeft als gevolg dat B hoger zal moeten zijn om deze maximale modelperformantie te kunnen bereiken. Bij de keuze van λ wordt dan ook vaak een afweging gemaakt tussen enerzijds λ zo klein mogelijk kiezen met het oog op een maximale modelperformantie en anderzijds λ groot genoeg kiezen, zodat de benodigde rekentijd om deze maximale modelperformantie te behalen, binnen aanvaardbare grenzen blijft.

BIJLAGE B

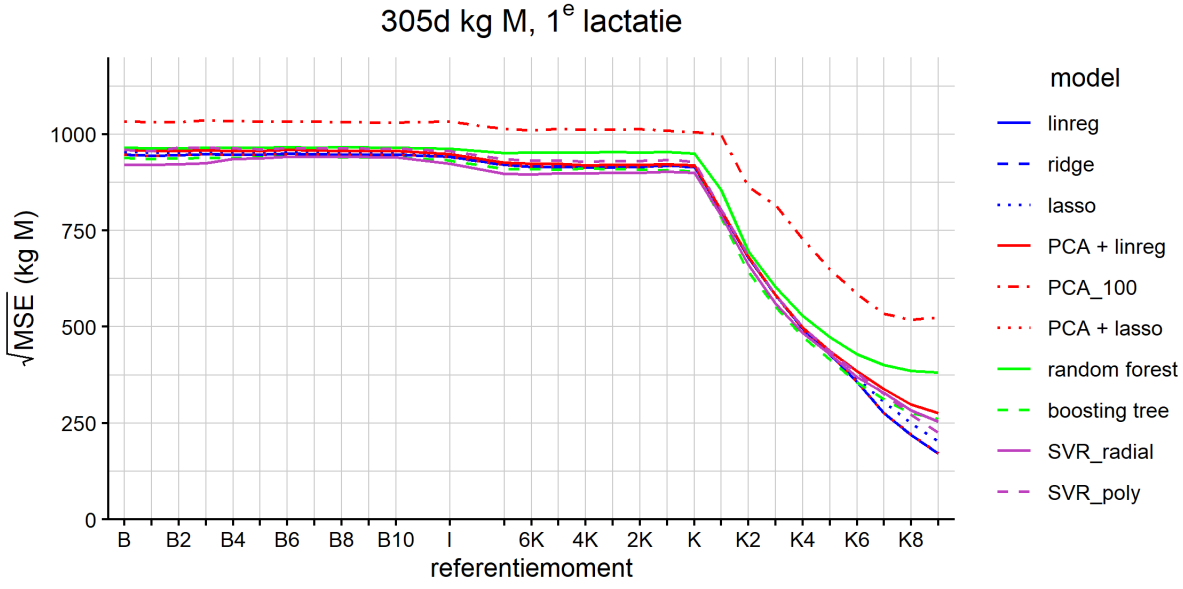
MODELRESULTATEN: FIGUREN

B.1 EVA: één model op landelijk niveau

- **305d kg M, 1e lactatie**

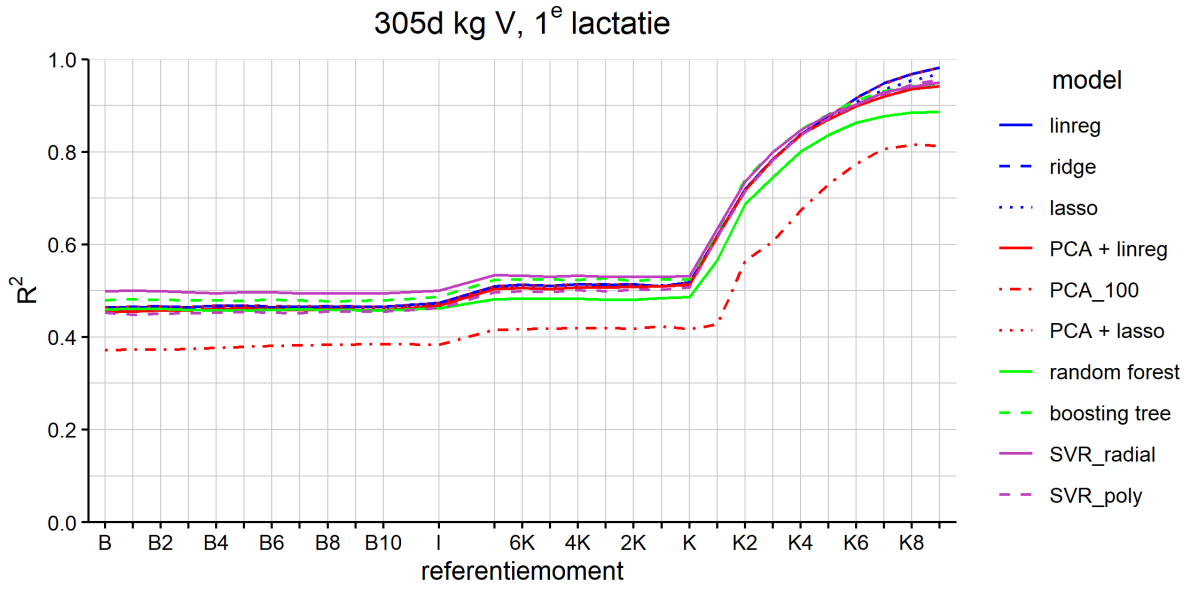


Figuur B.1: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg melk

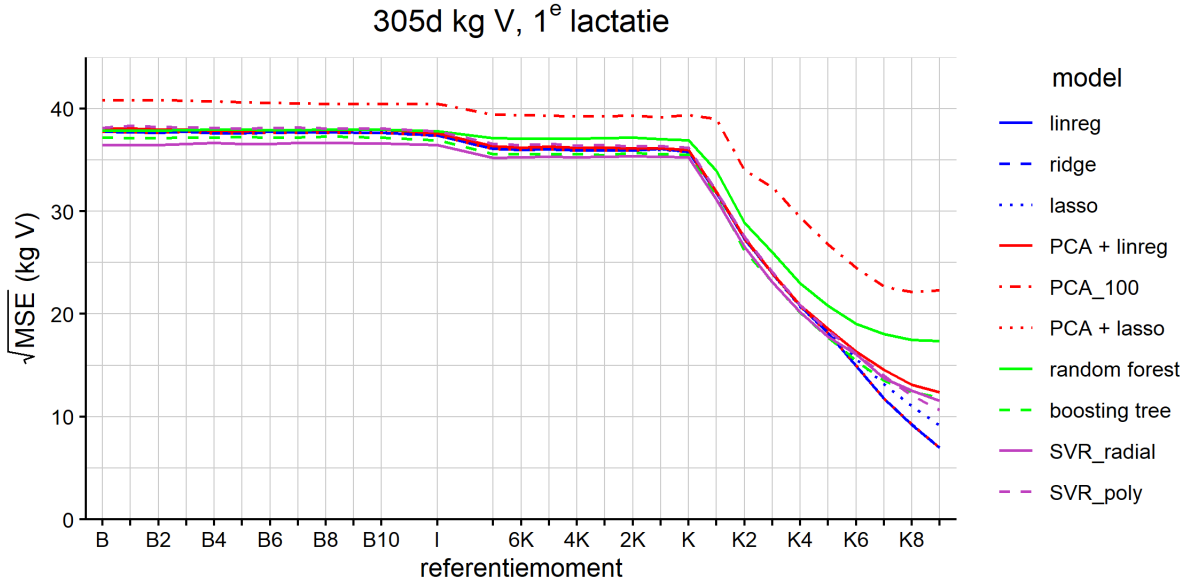


Figuur B.2: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg melk

- **305d kg V, 1e lactatie**

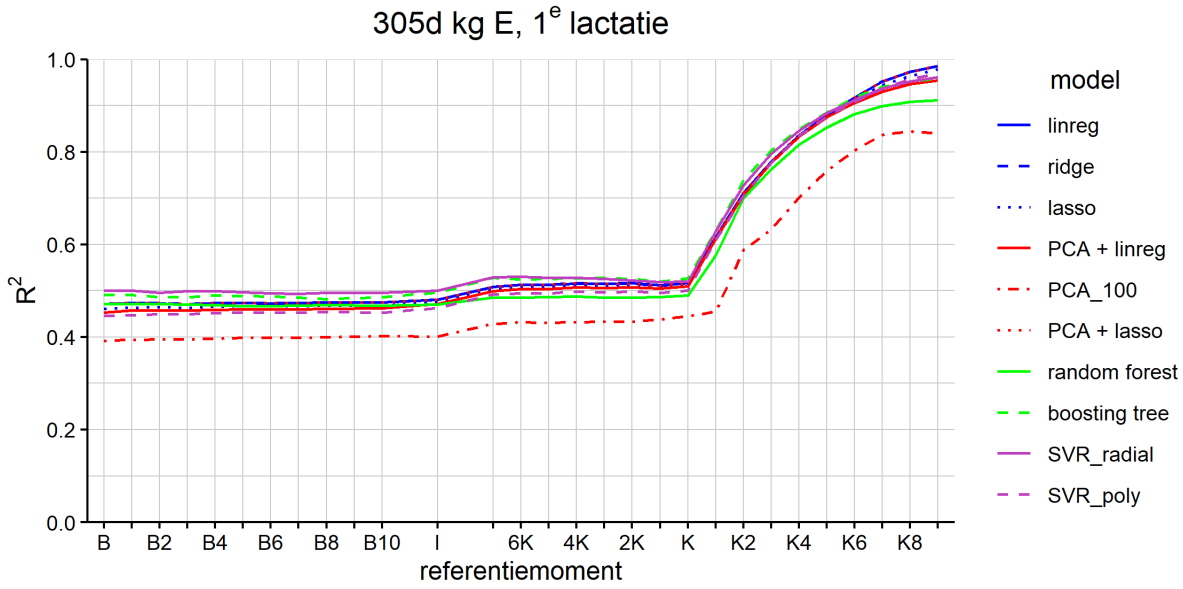


Figuur B.3: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg vet

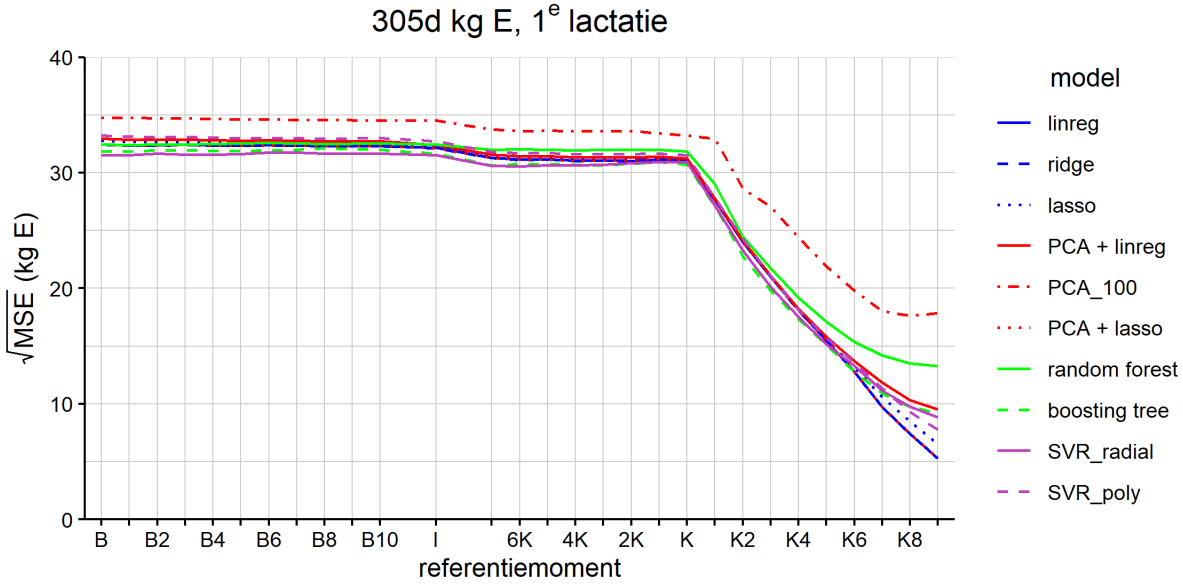


Figuur B.4: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg vet

- **305d kg E, 1e lactatie**

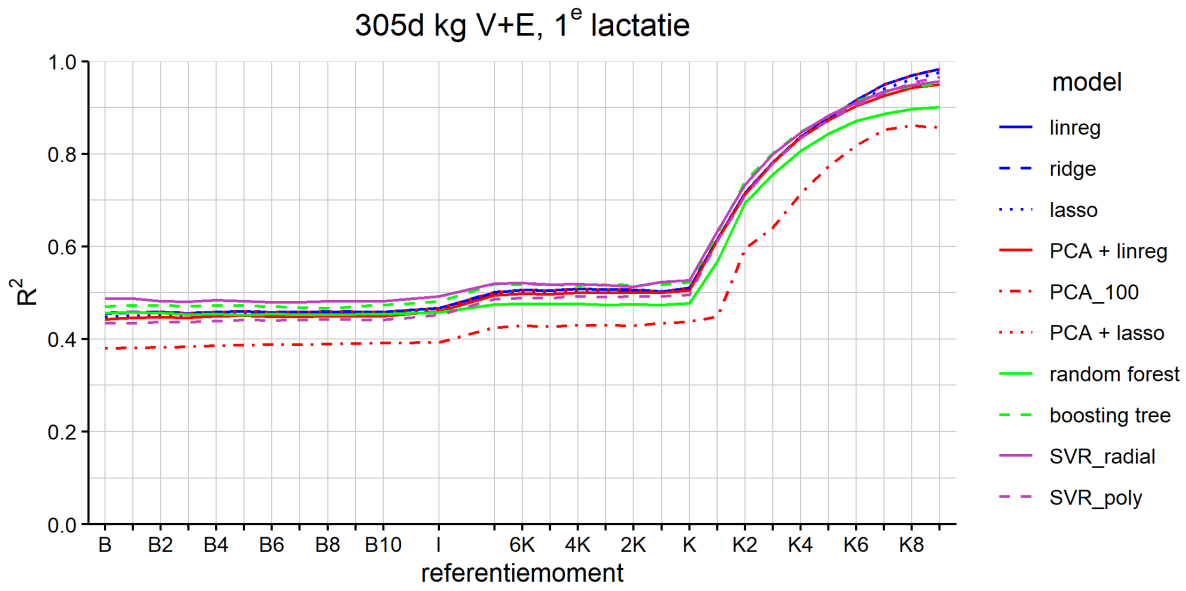


Figuur B.5: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg eiwit

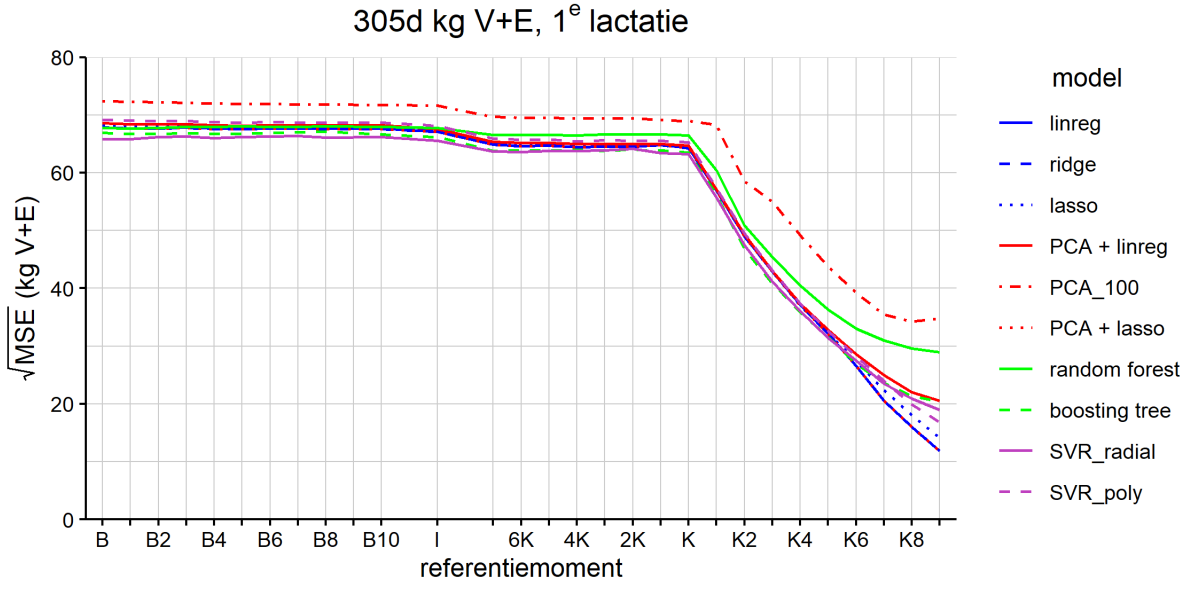


Figuur B.6: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg eiwit

- **305d kg V+E, 1e lactatie**

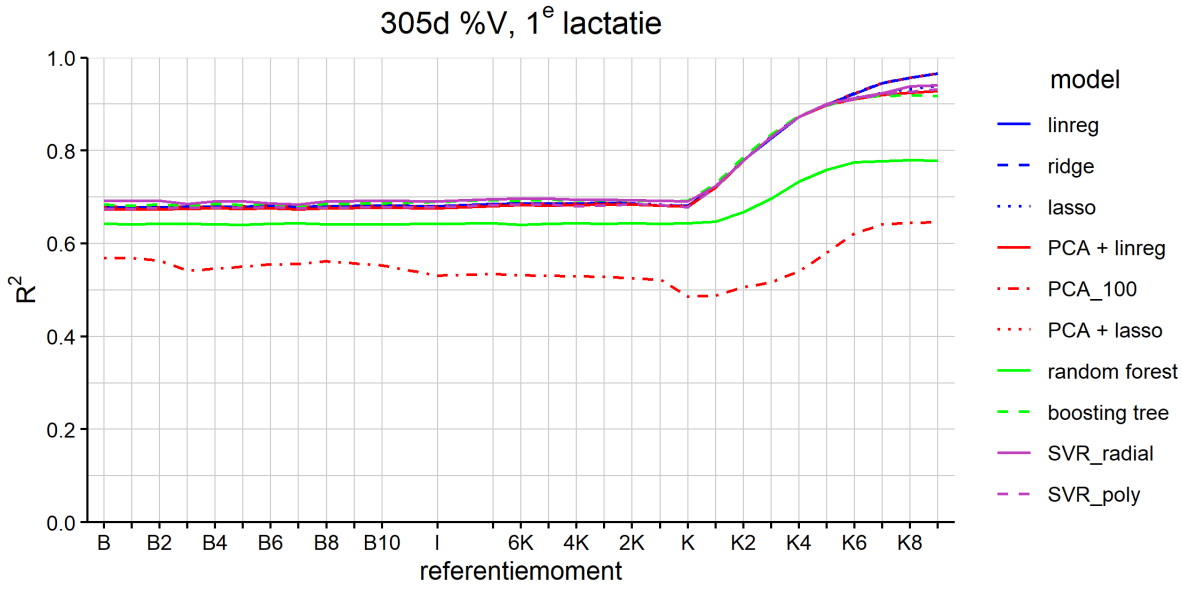


Figuur B.7: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg vet en eiwit

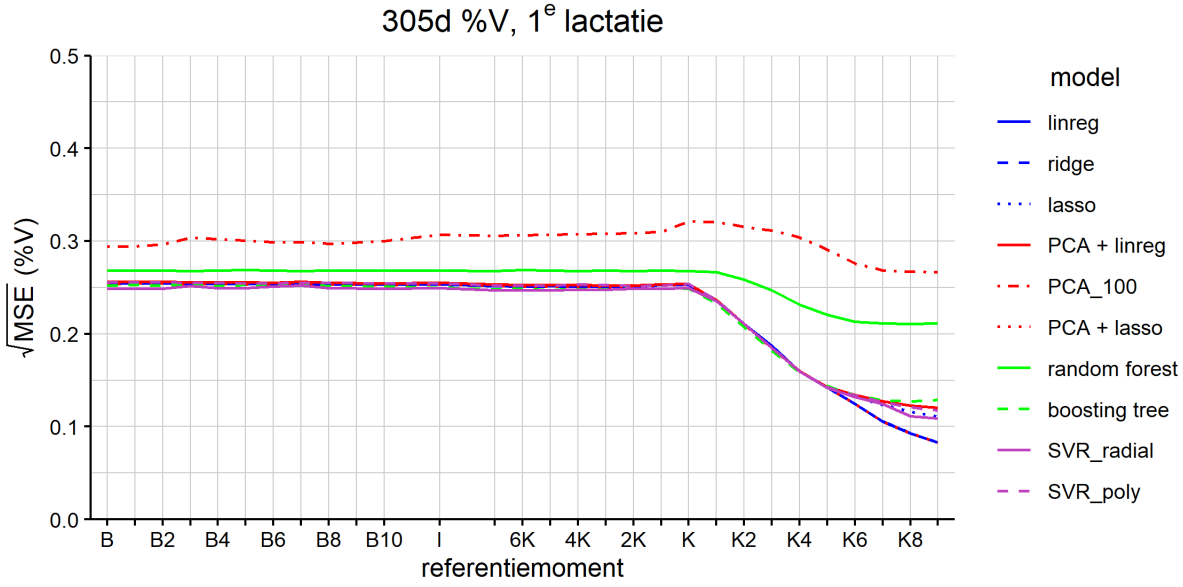


Figuur B.8: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg vet en eiwit

- **305d %V, 1e lactatie**

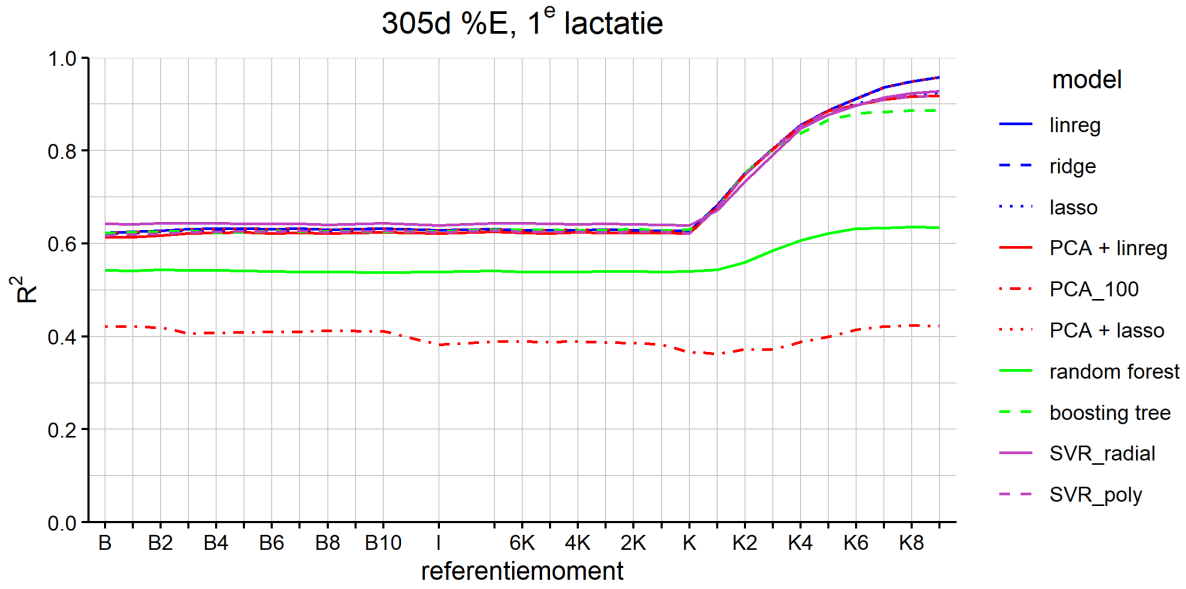


Figuur B.9: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in procent vet

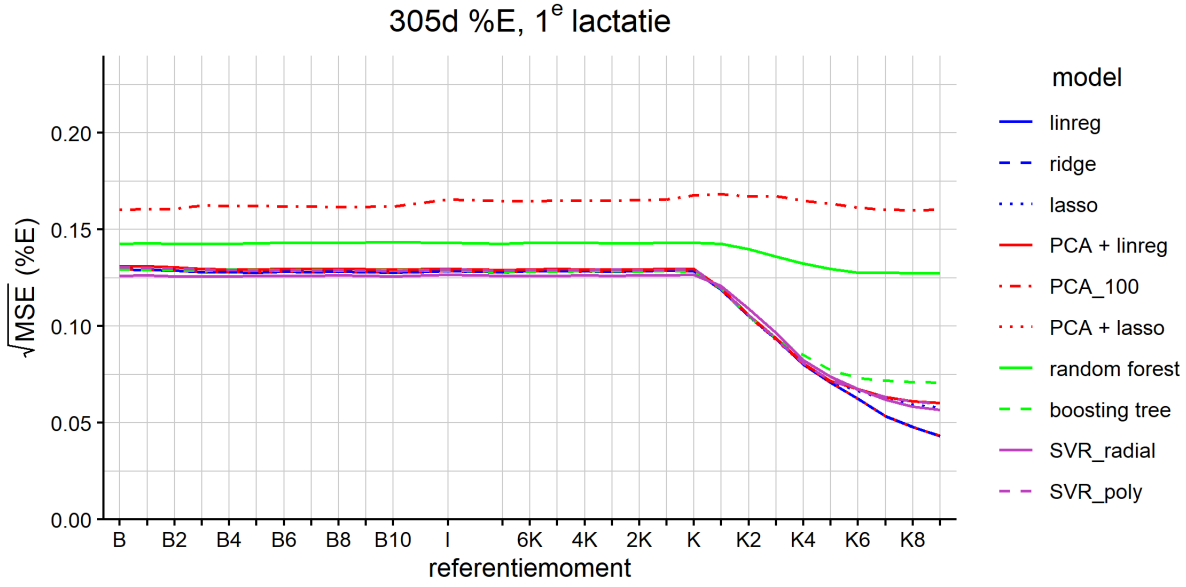


Figuur B.10: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in procent vet

- **305d %E, 1e lactatie**

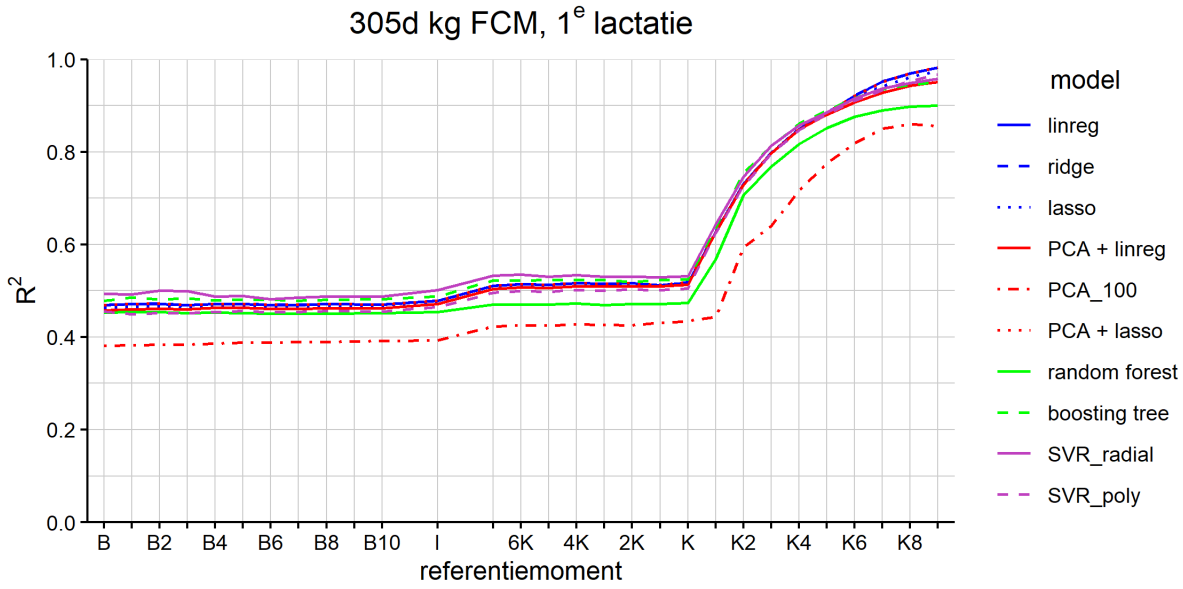


Figuur B.11: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in procent eiwit

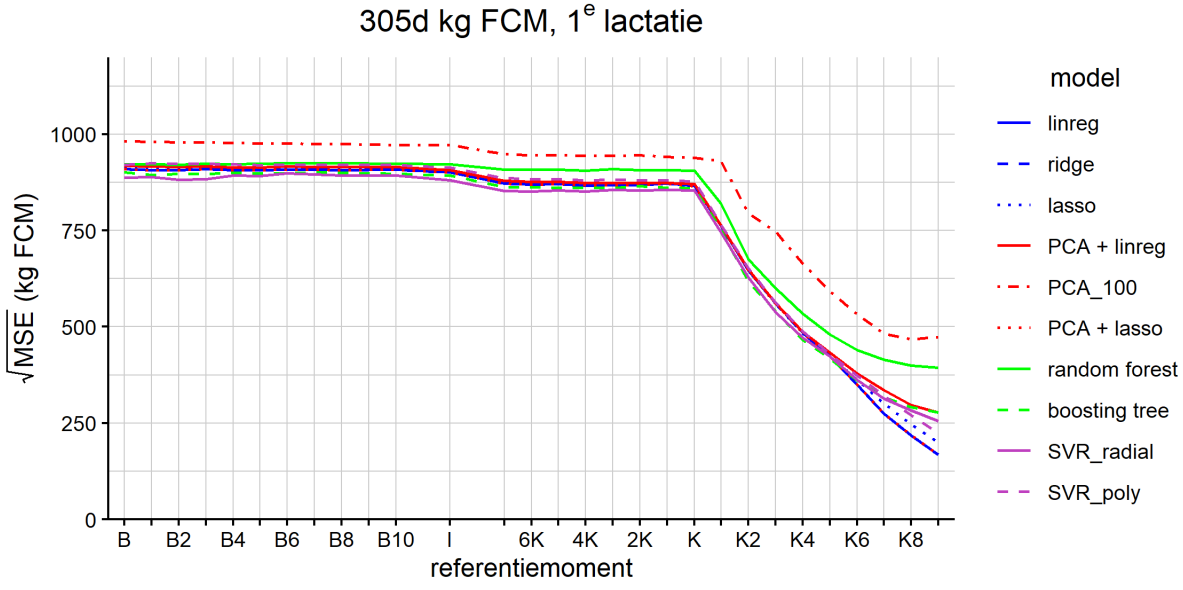


Figuur B.12: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in procent eiwit

- **305d kg FCM, 1e lactatie**

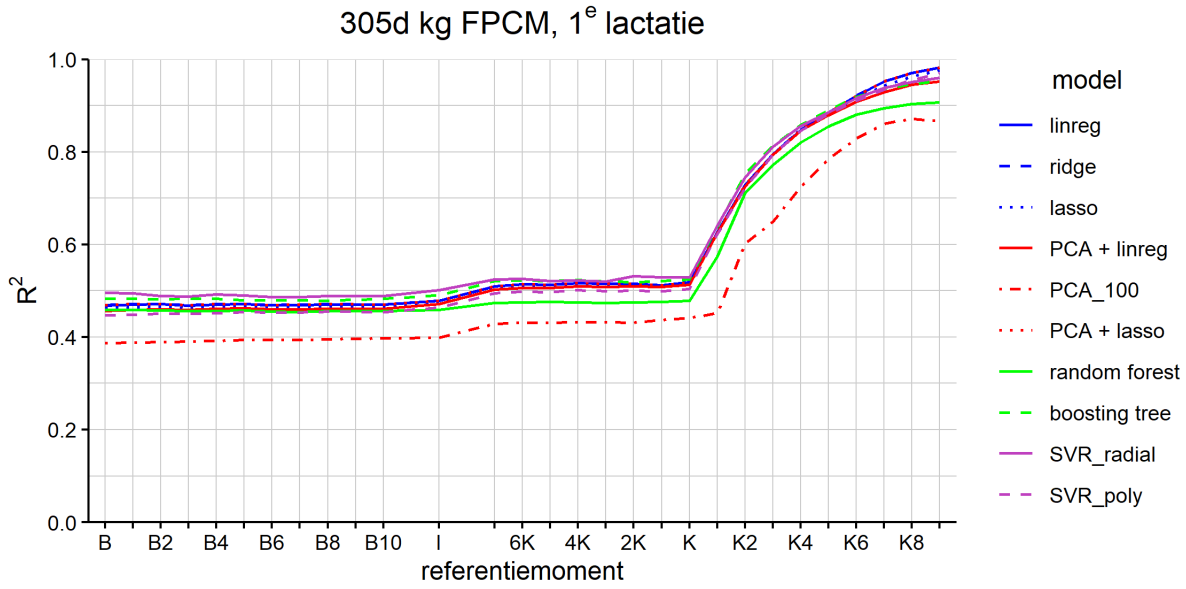


Figuur B.13: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg FCM

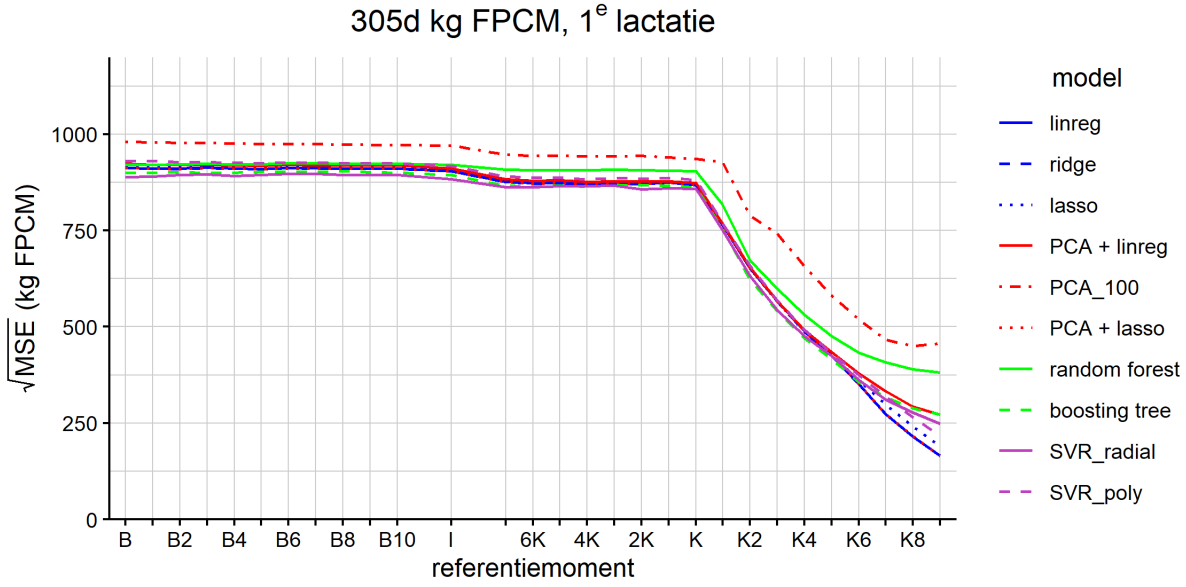


Figuur B.14: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg FCM

- **305d kg FPCM, 1e lactatie**

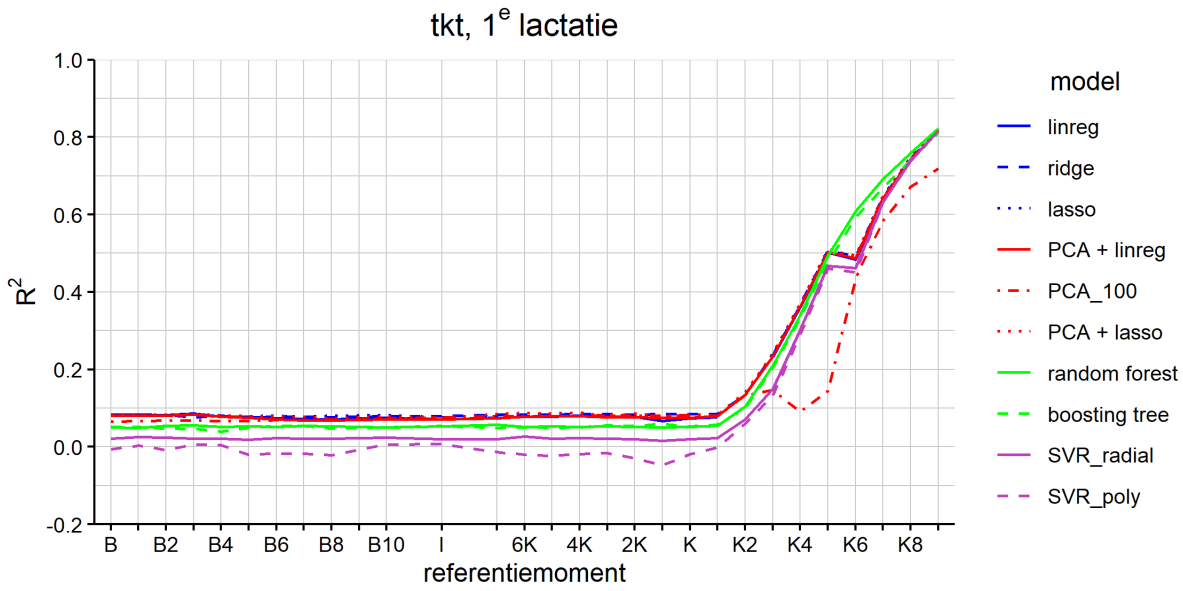


Figuur B.15: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg FPCM

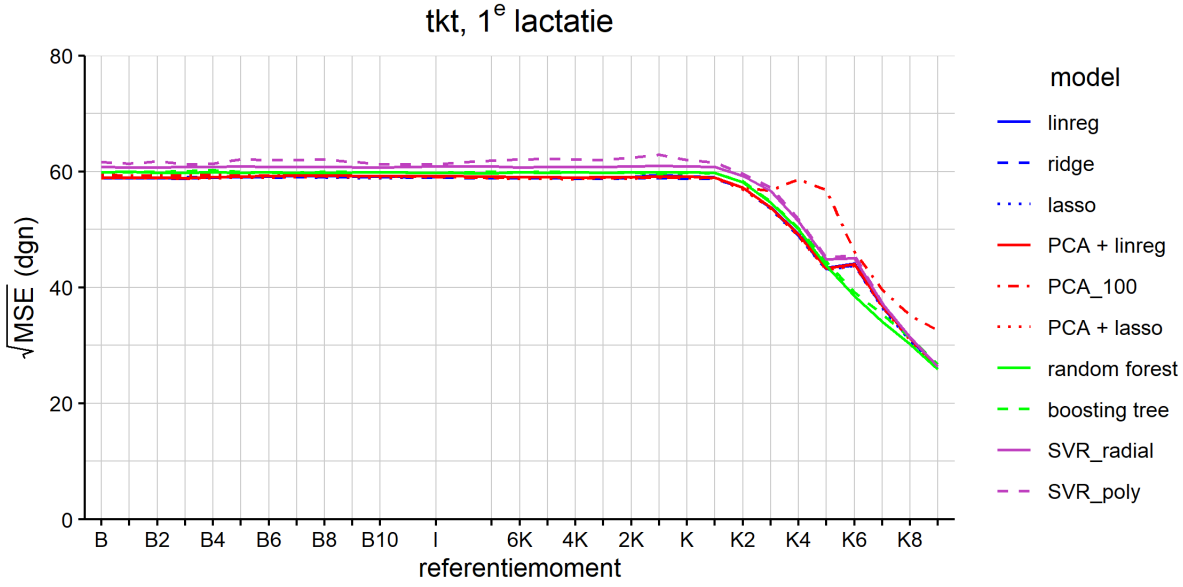


Figuur B.16: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg FPCM

- **tkt, 1e lactatie**

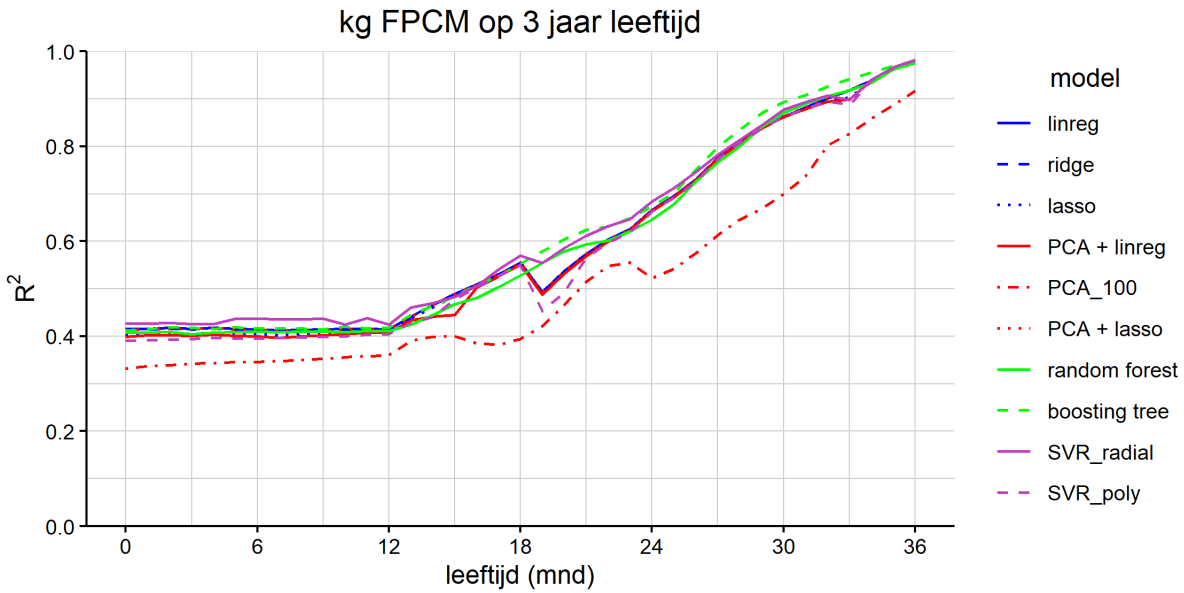


Figuur B.17: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de tussenkalftijd tussen de eerste en de tweede pariteit

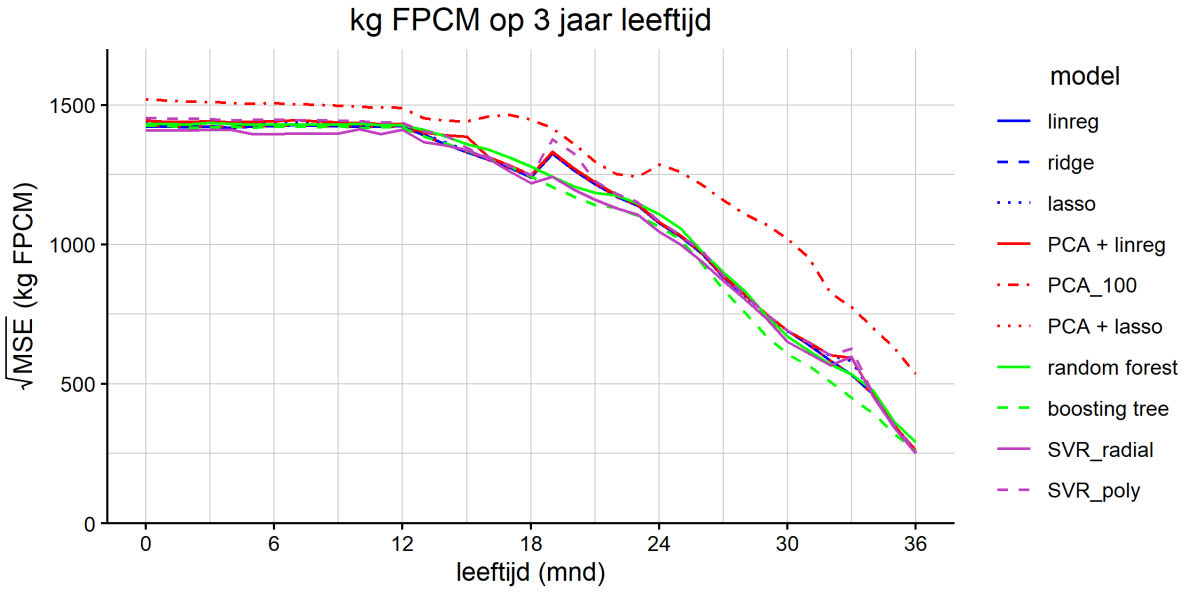


Figuur B.18: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de tussenkalftijd tussen de eerste en de tweede pariteit

- **kg FPCM op 3 jaar leeftijd**

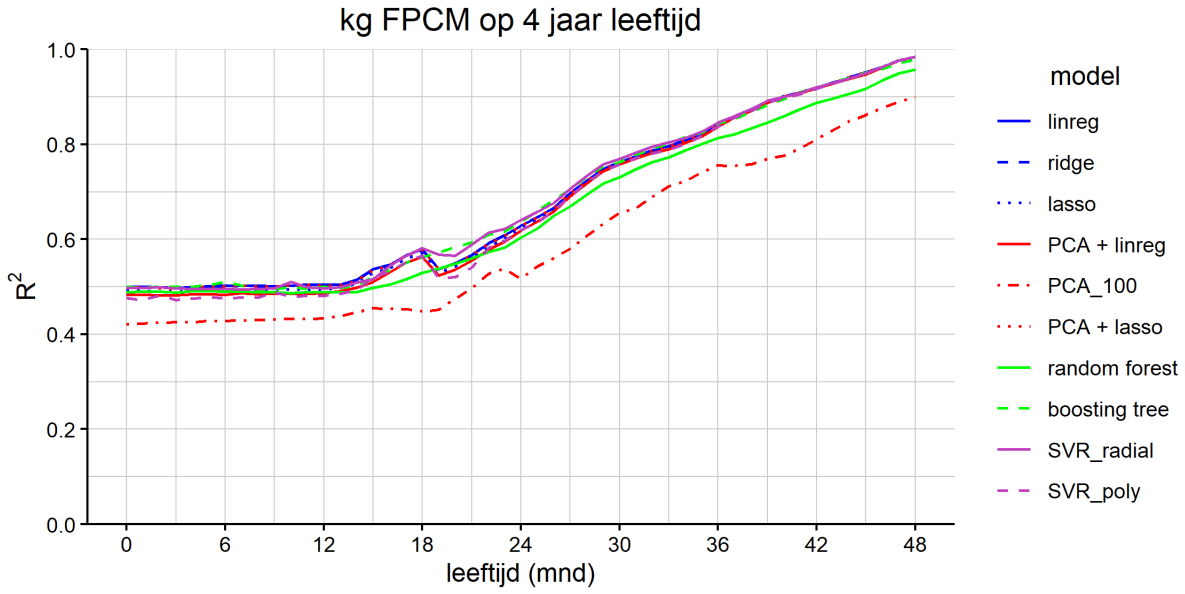


Figuur B.19: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 3 jaar

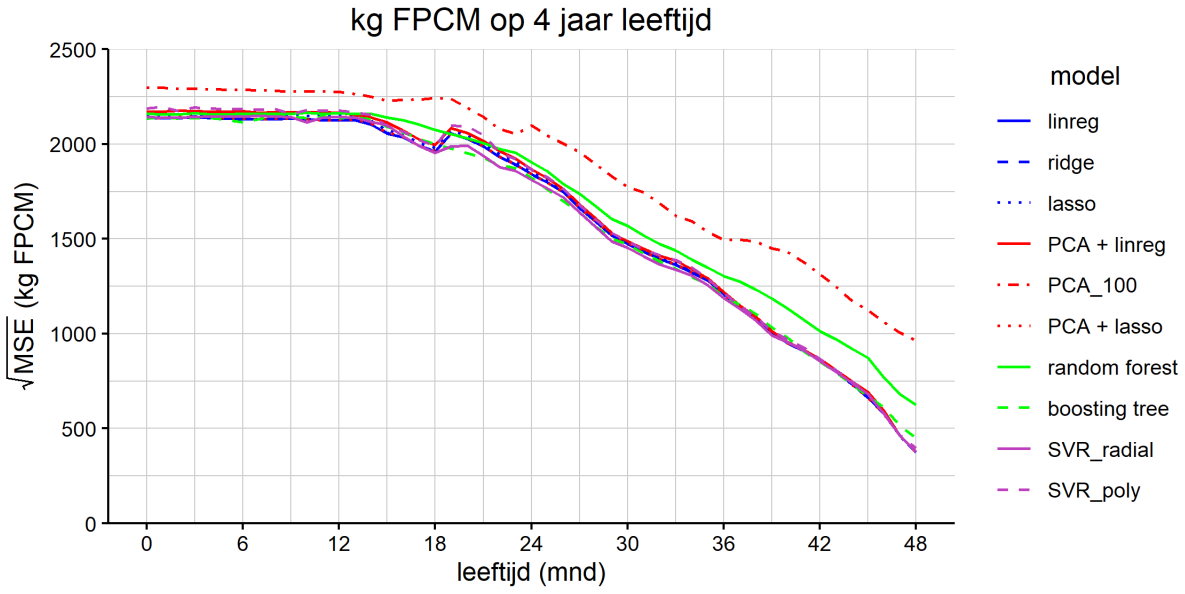


Figuur B.20: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 3 jaar

- **kg FPCM op 4 jaar leeftijd**

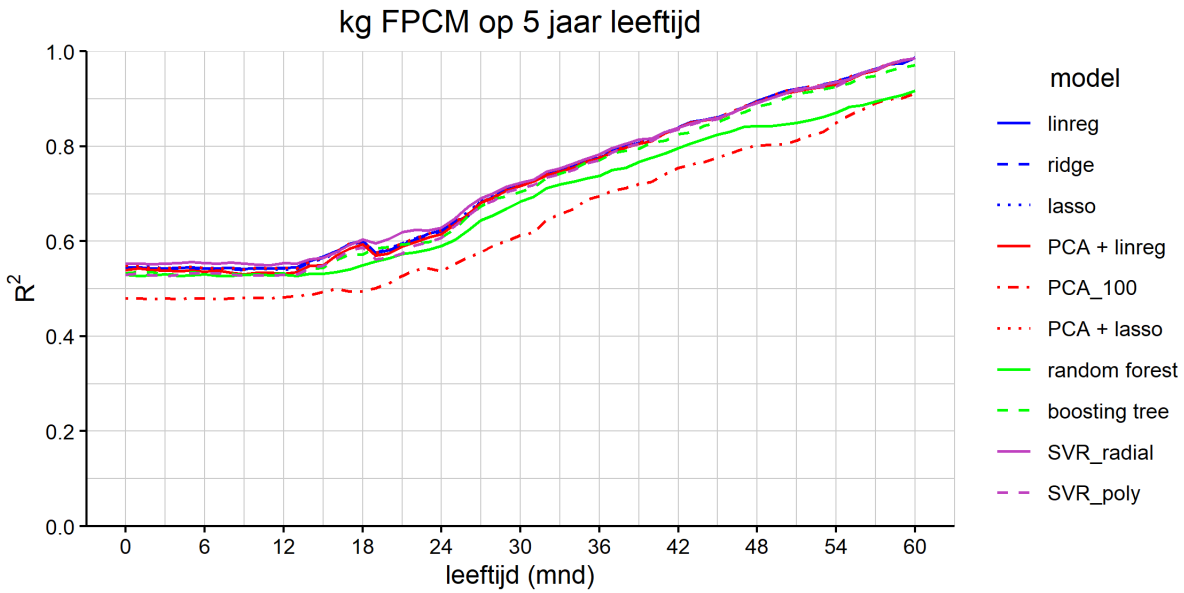


Figuur B.21: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 4 jaar

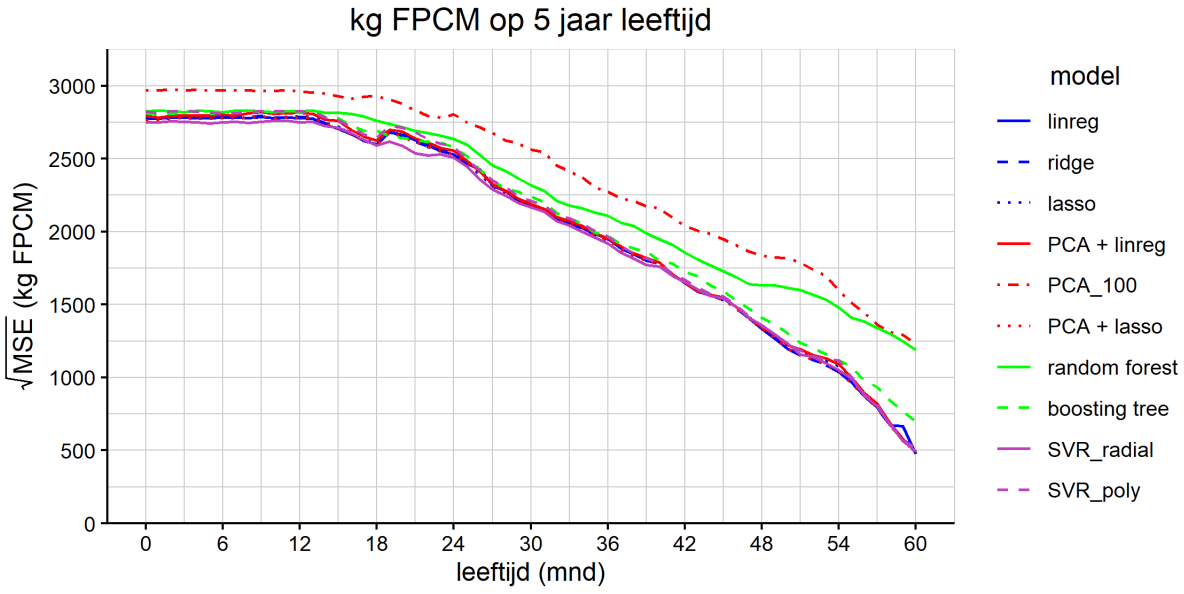


Figuur B.22: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 4 jaar

- **kg FPCM op 5 jaar leeftijd**

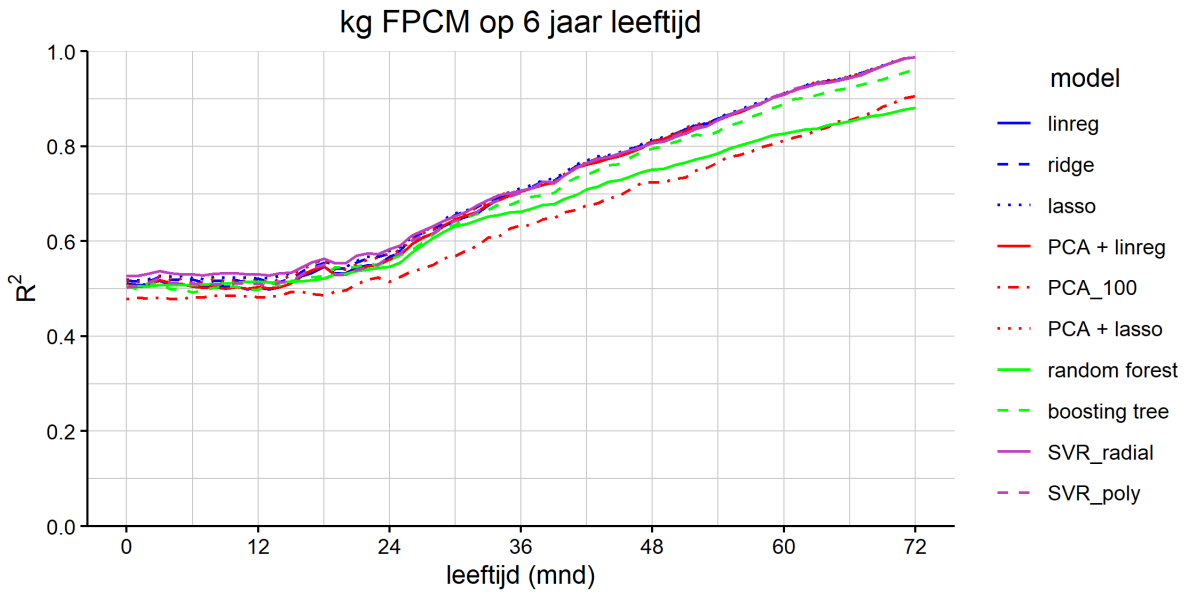


Figuur B.23: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 5 jaar

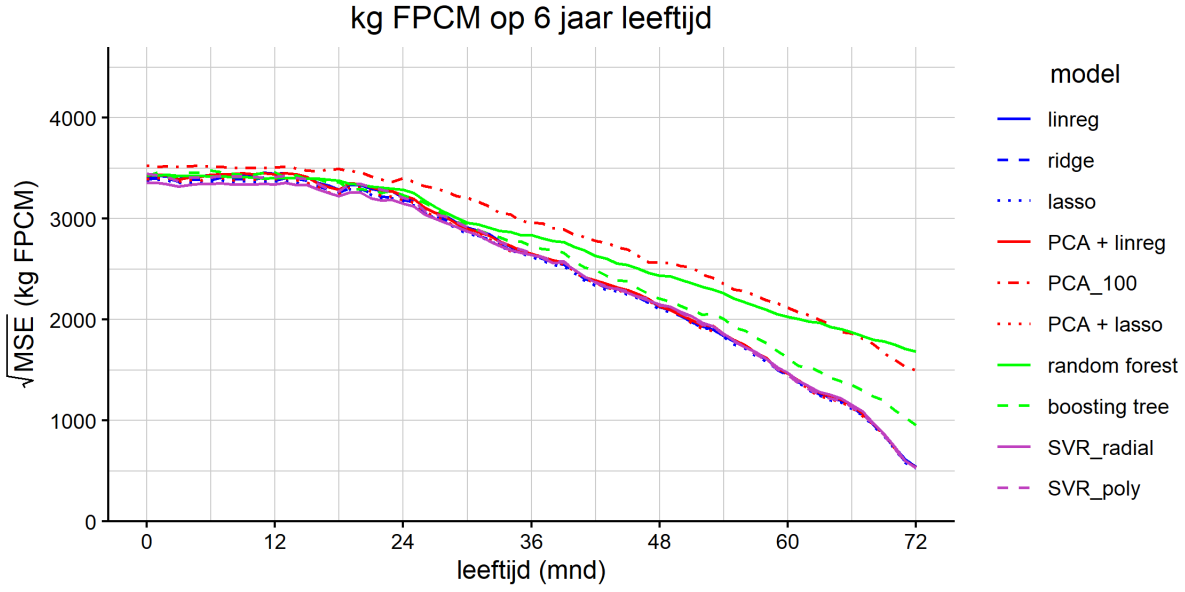


Figuur B.24: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 5 jaar

- **kg FPCM op 6 jaar leeftijd**

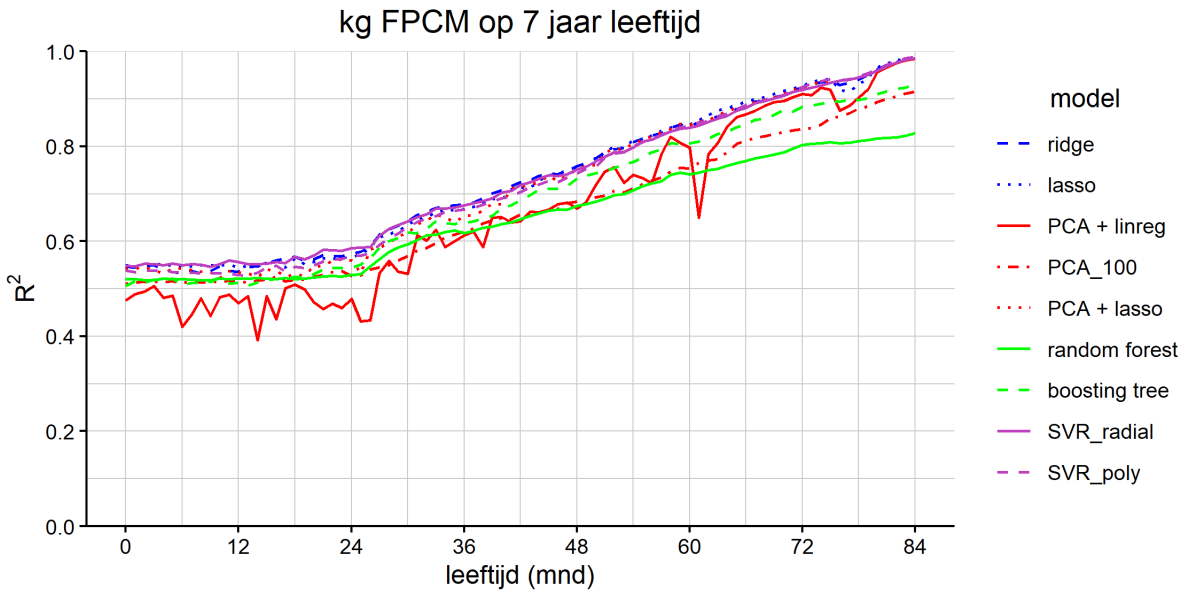


Figuur B.25: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 6 jaar

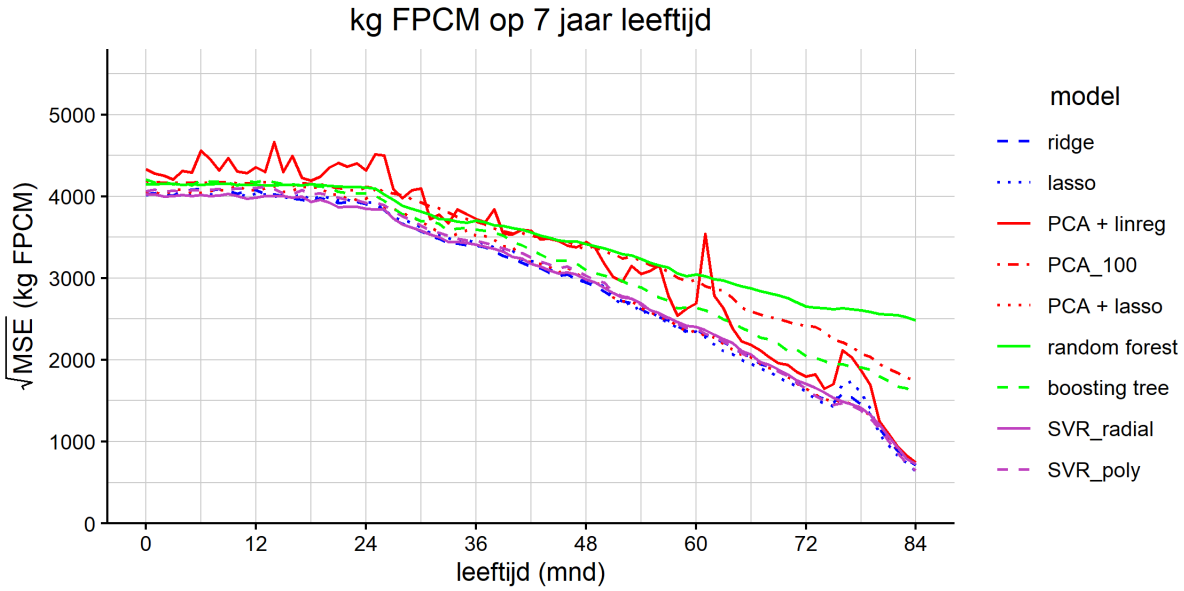


Figuur B.26: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 6 jaar

- **kg FPCM op 7 jaar leeftijd**

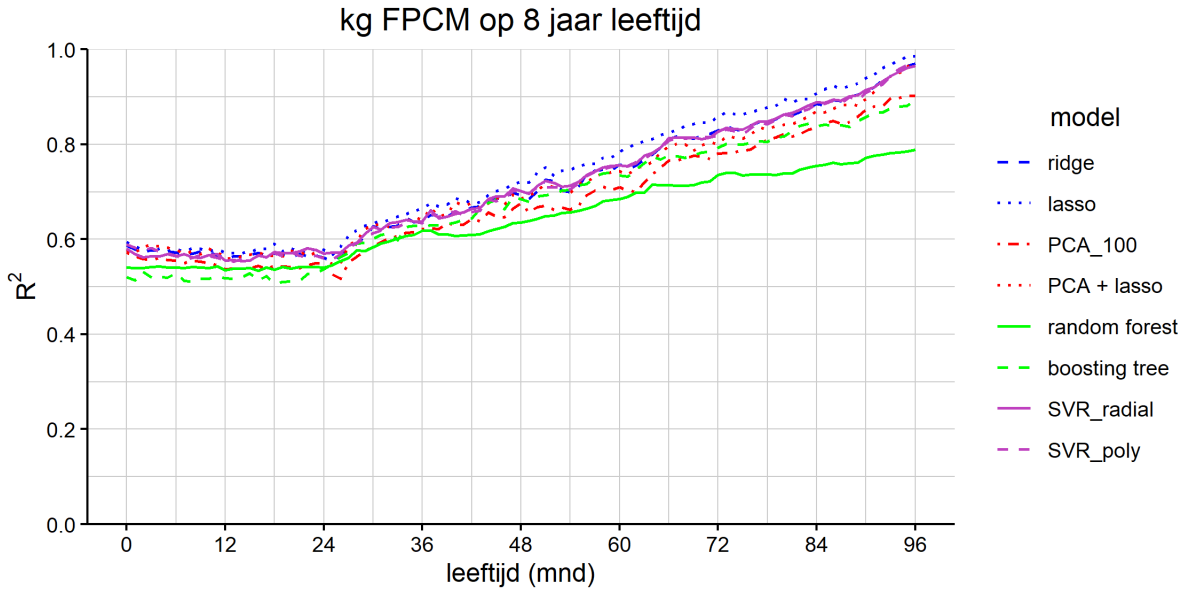


Figuur B.27: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 7 jaar

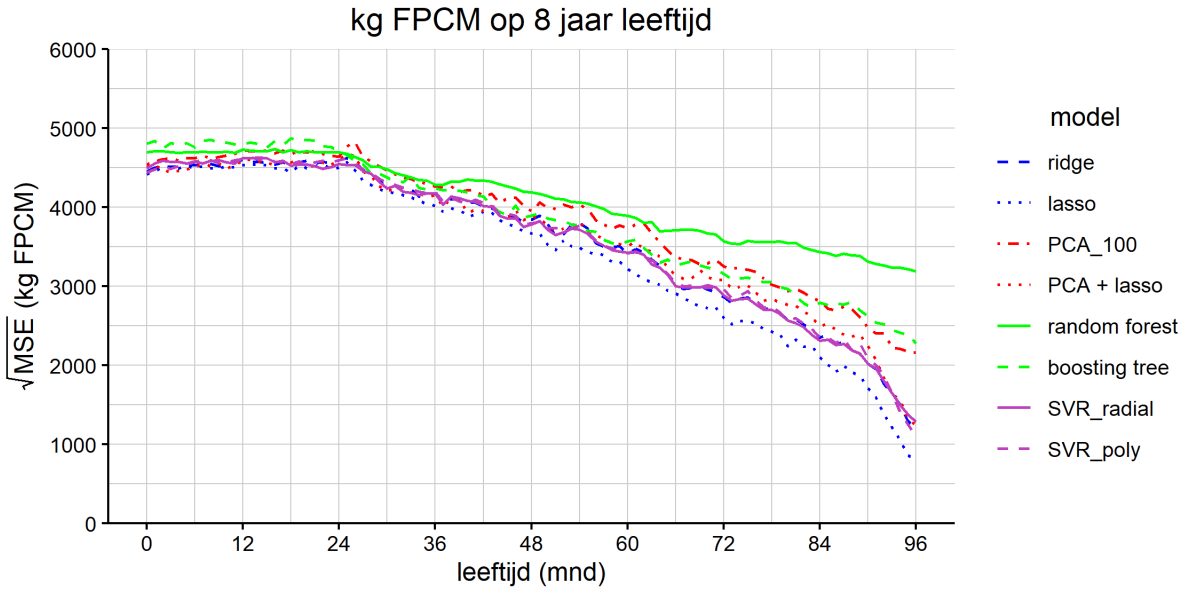


Figuur B.28: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 7 jaar

- **kg FPCM op 8 jaar leeftijd**

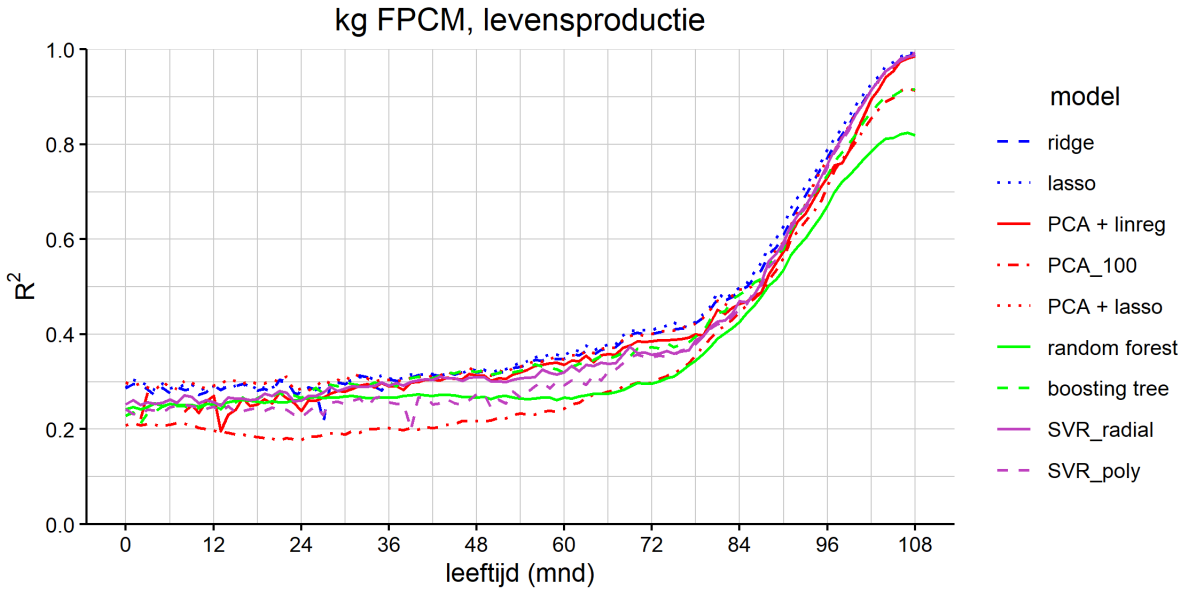


Figuur B.29: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 8 jaar

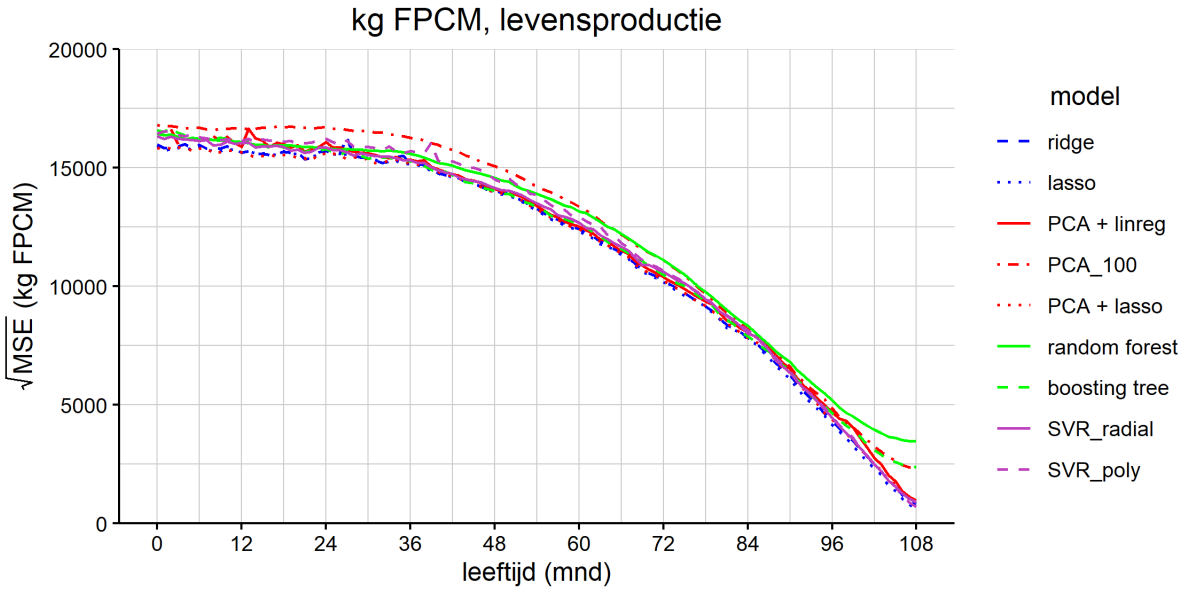


Figuur B.30: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 8 jaar

- **kg FPCM, levensproductie**



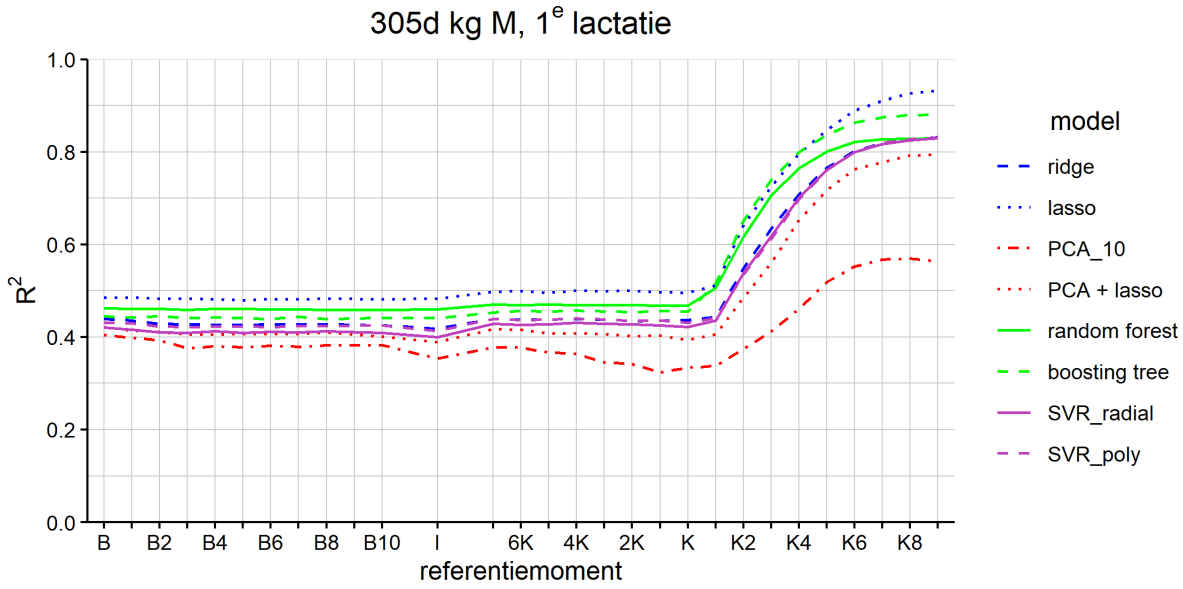
Figuur B.31: R^2-waarden op landelijk niveau voor de EVA-modellen met betrekking tot de levensproductie, uitgedrukt in kg FPCM



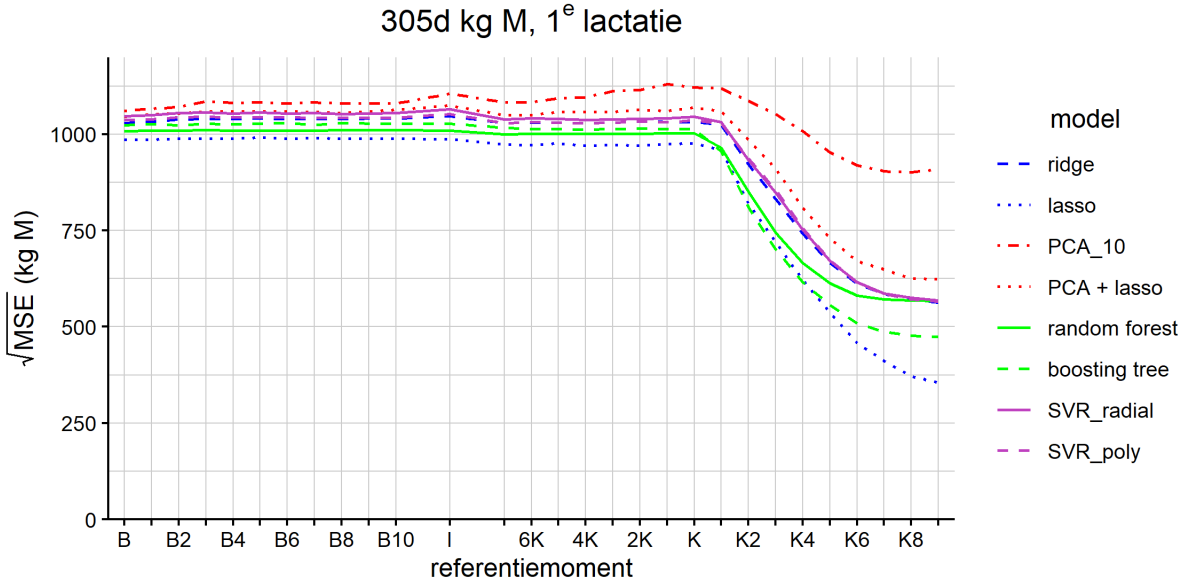
Figuur B.32: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVA-modellen met betrekking tot de levensproductie, uitgedrukt in kg FPCM

B.2 EVI: één model voor ieder bedrijf

- **305d kg M, 1e lactatie**

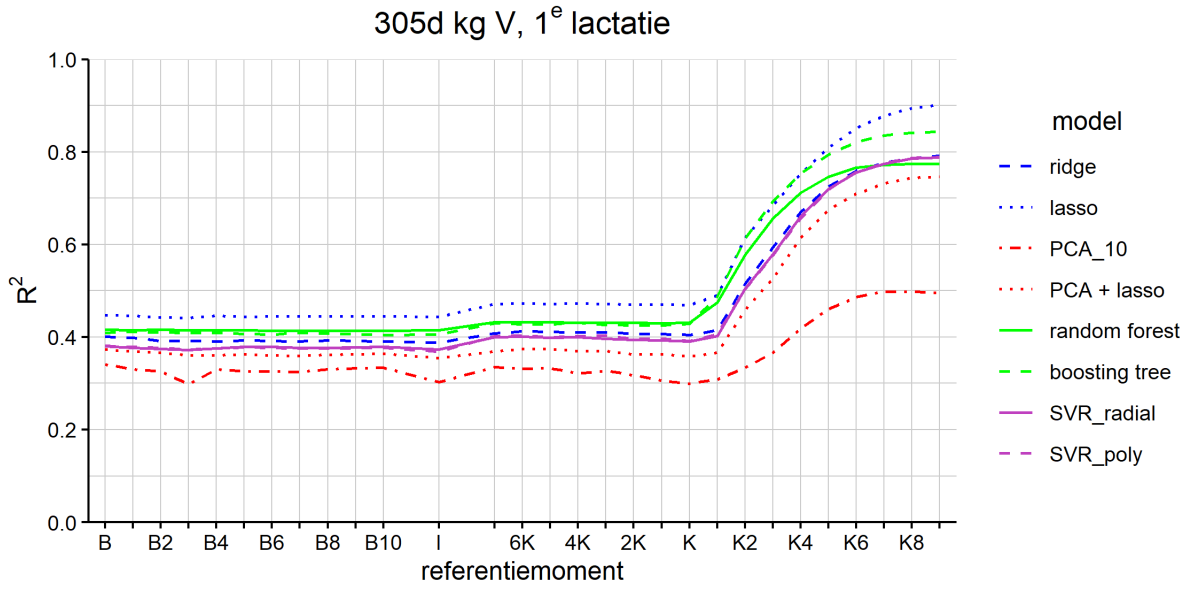


Figuur B.33: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg melk

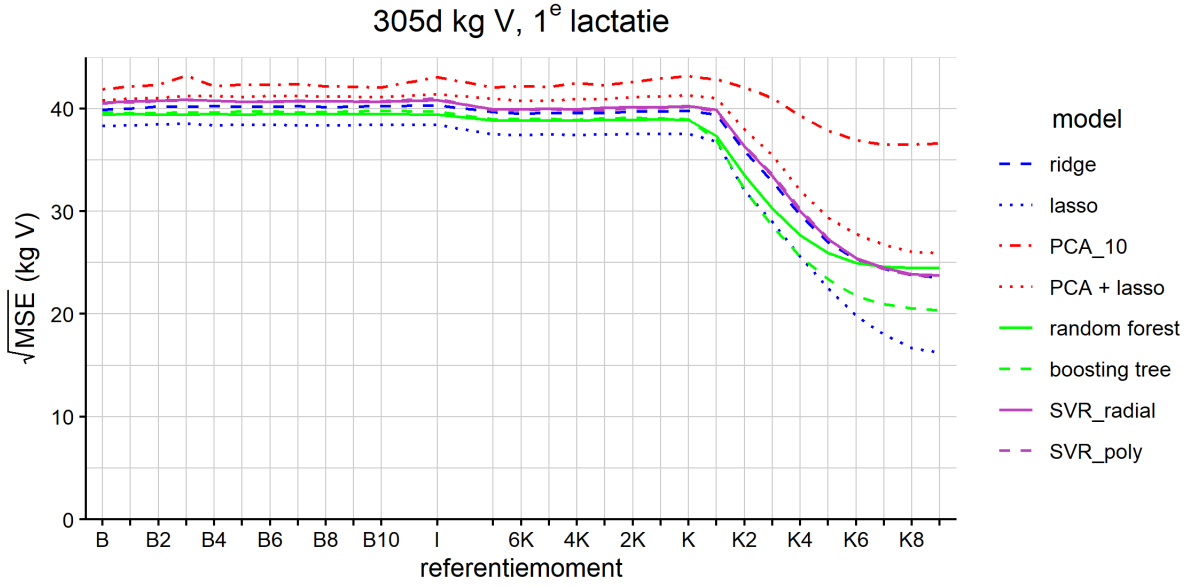


Figuur B.34: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg melk

- **305d kg V, 1e lactatie**

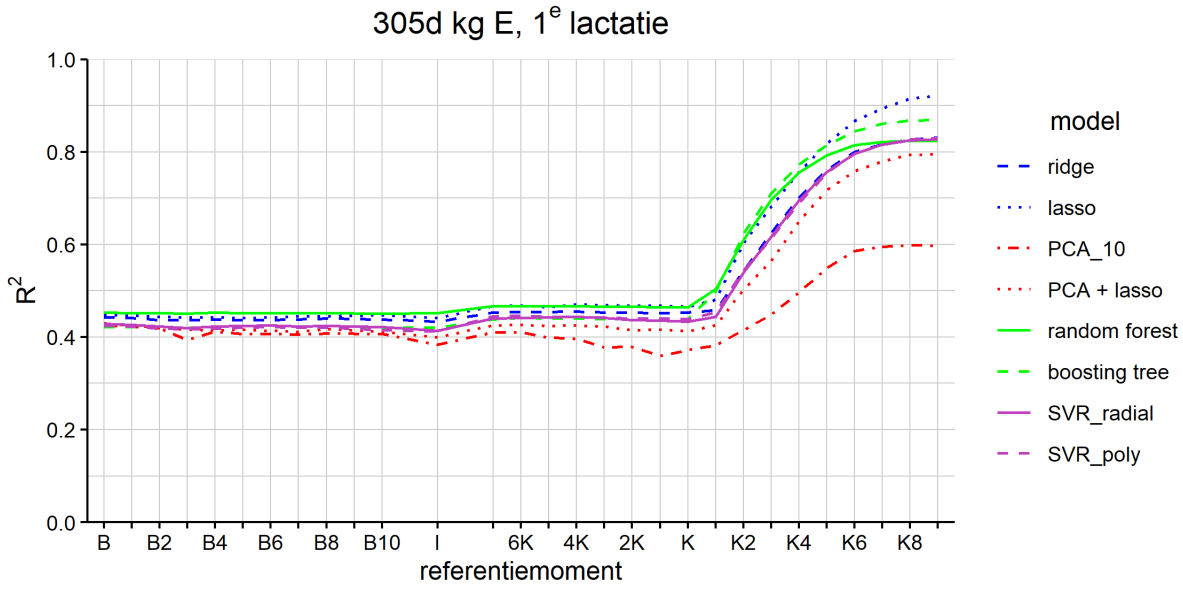


Figuur B.35: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg vet

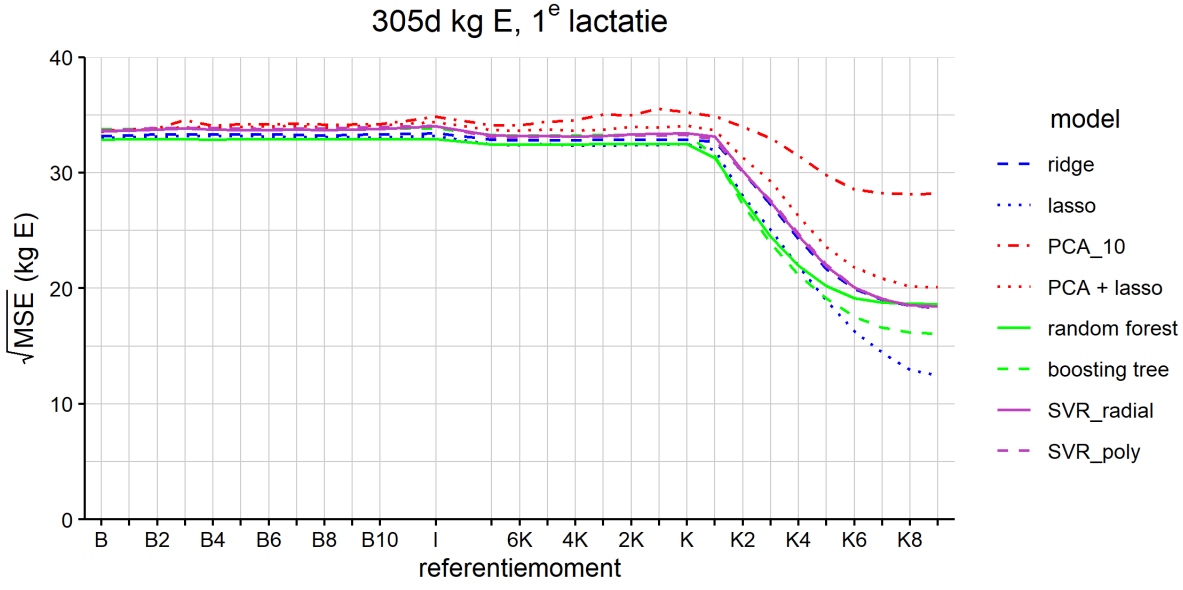


Figuur B.36: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg vet

- **305d kg E, 1e lactatie**

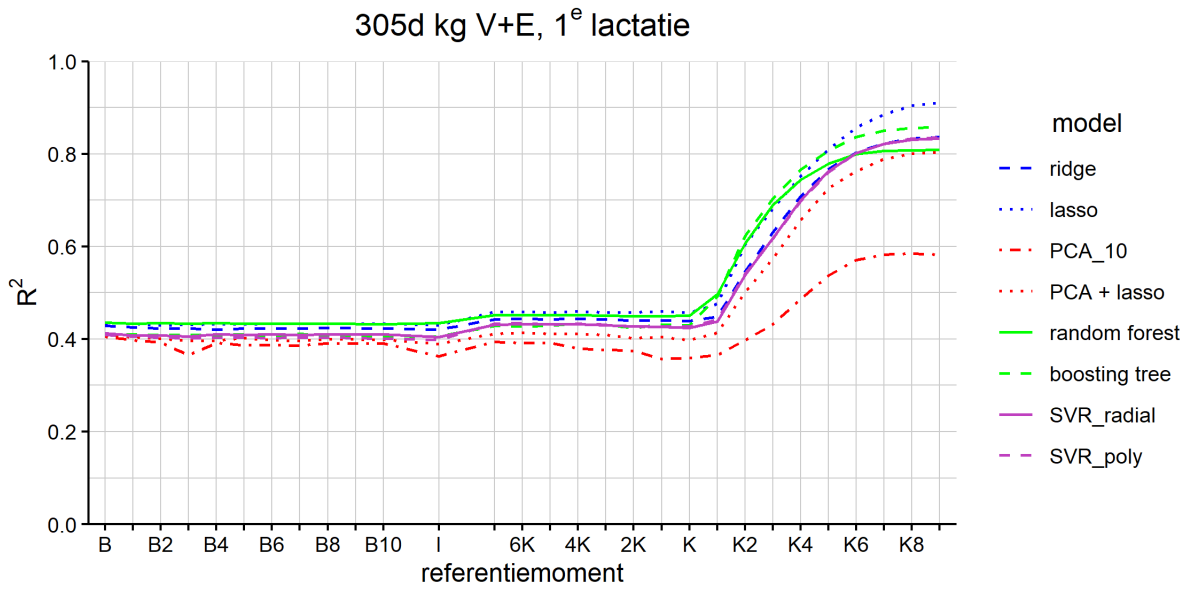


Figuur B.37: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg eiwit

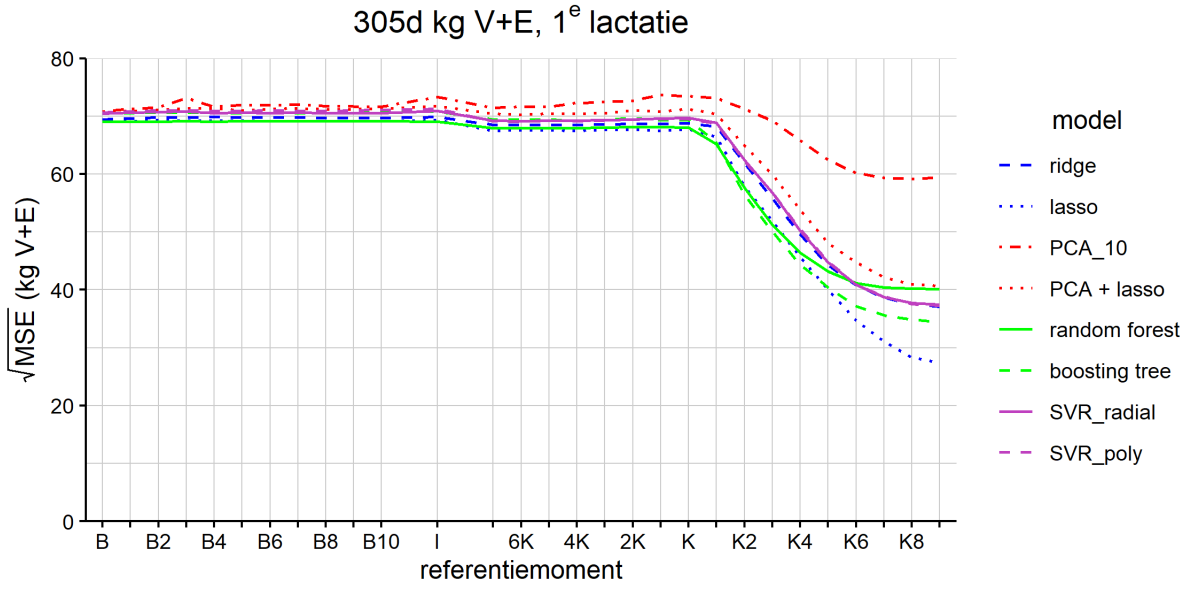


Figuur B.38: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg eiwit

- **305d kg V+E, 1e lactatie**

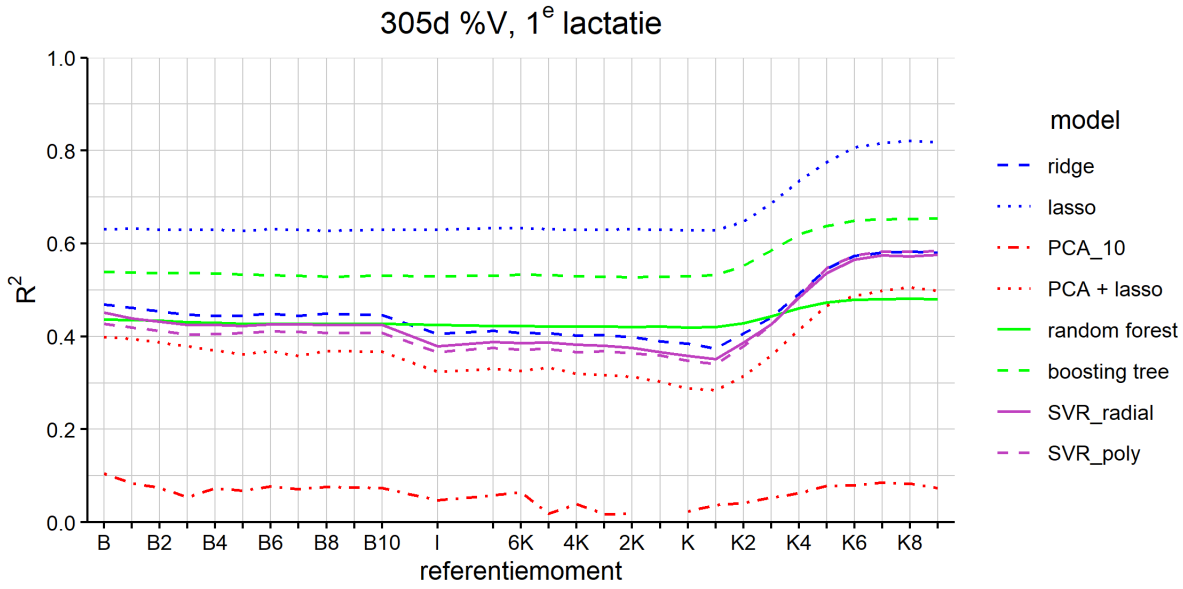


Figuur B.39: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg vet en eiwit

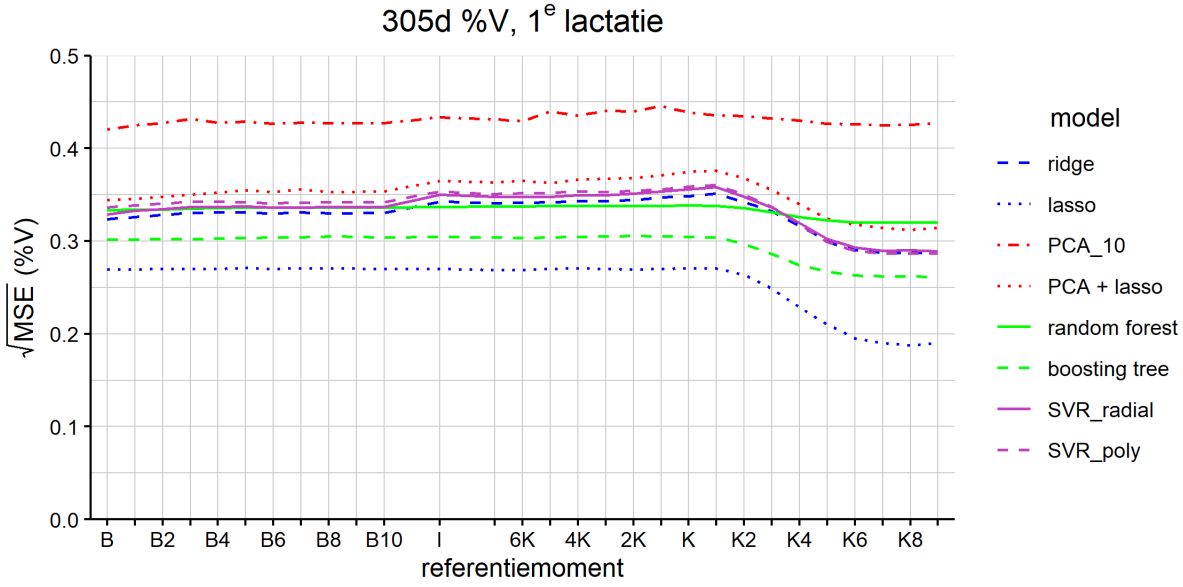


Figuur B.40: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg vet en eiwit

- **305d %V, 1e lactatie**

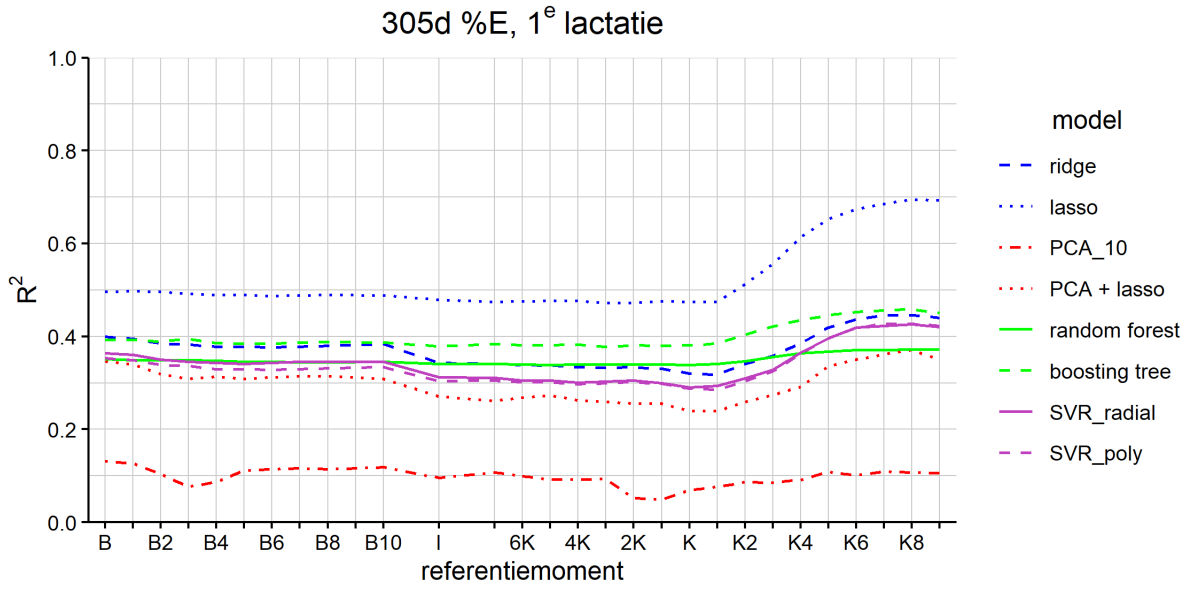


Figuur B.41: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in procent vet

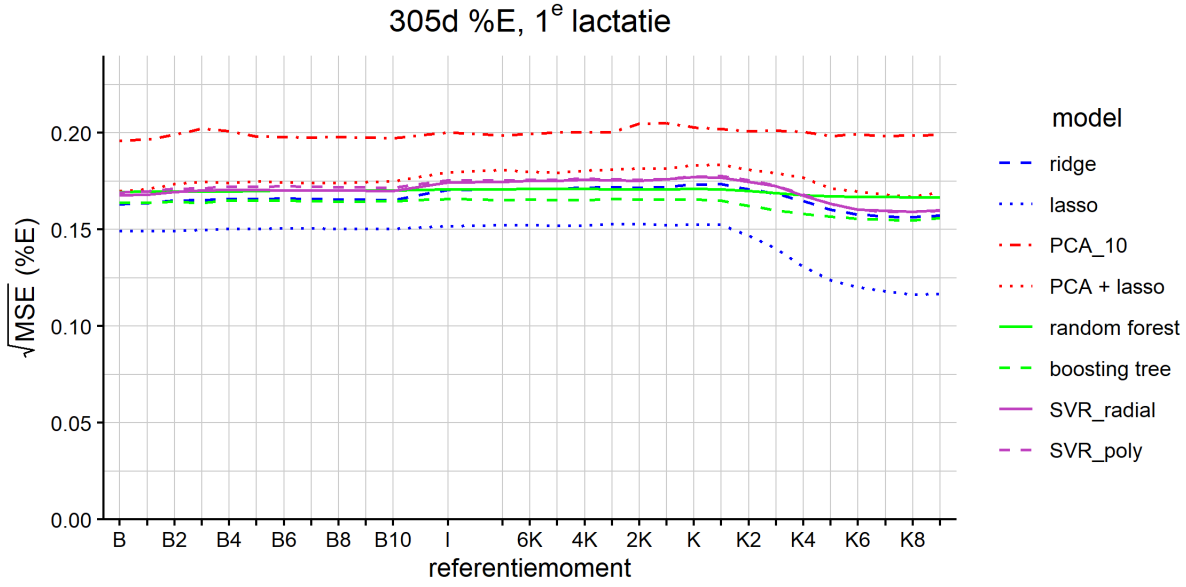


Figuur B.42: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in procent vet

- **305d %E, 1e lactatie**

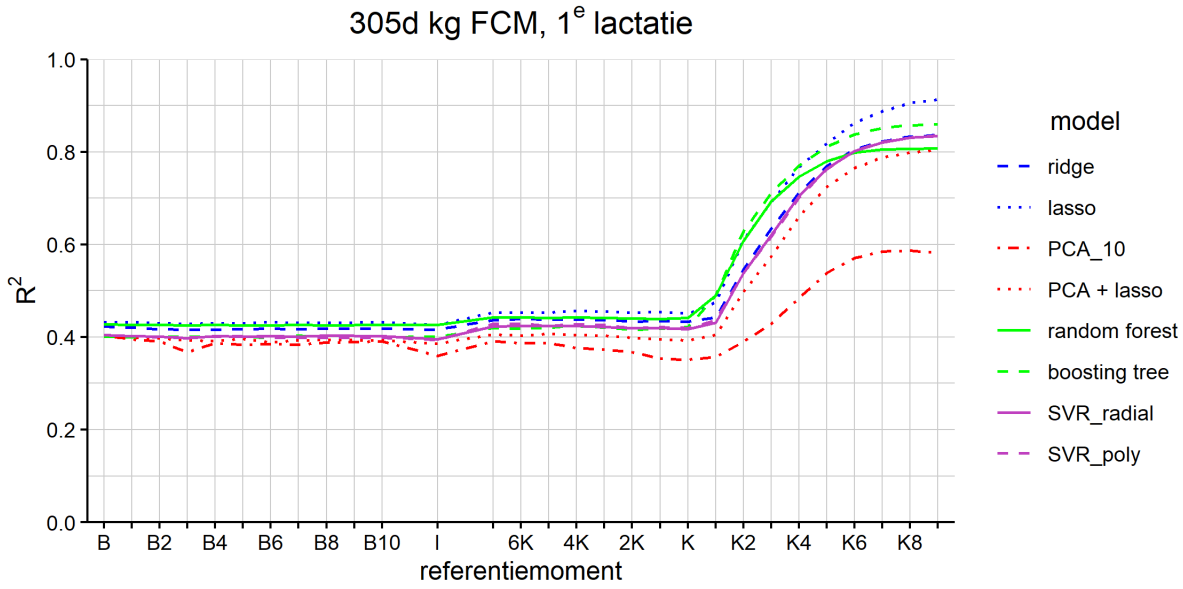


Figuur B.43: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in procent eiwit

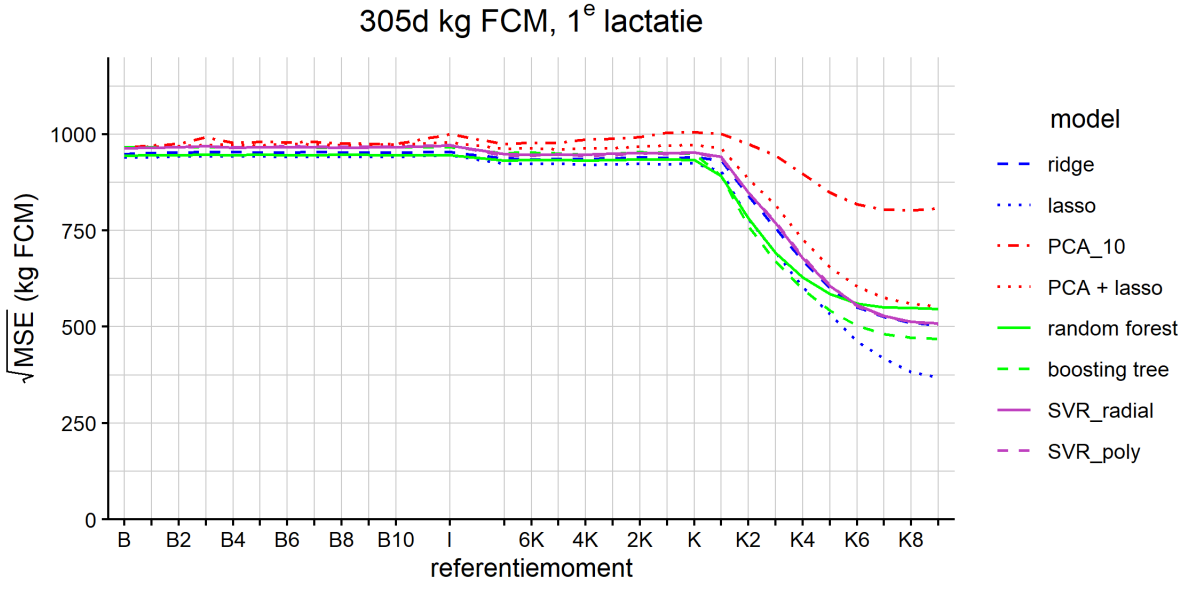


Figuur B.44: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in procent eiwit

- **305d kg FCM, 1e lactatie**

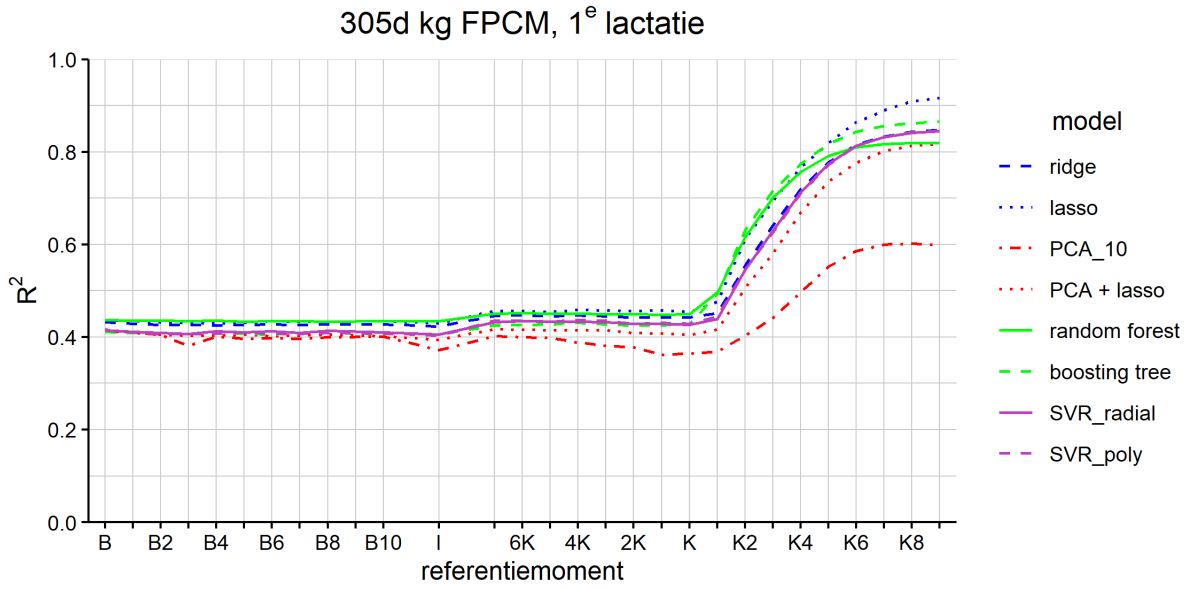


Figuur B.45: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg FCM

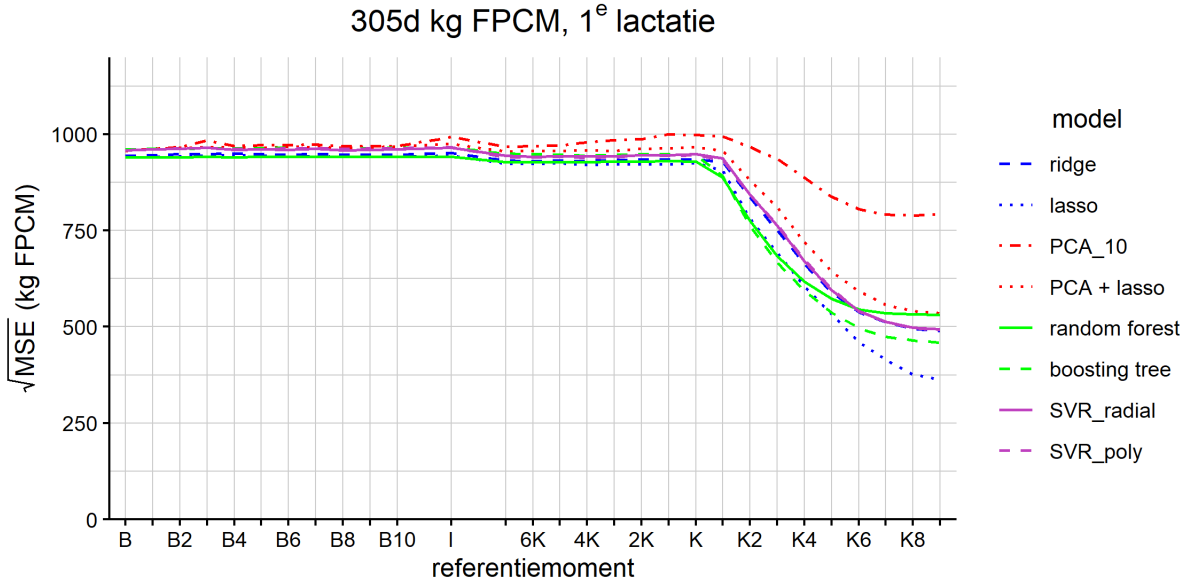


Figuur B.46: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg FCM

- **305d kg FPCM, 1e lactatie**

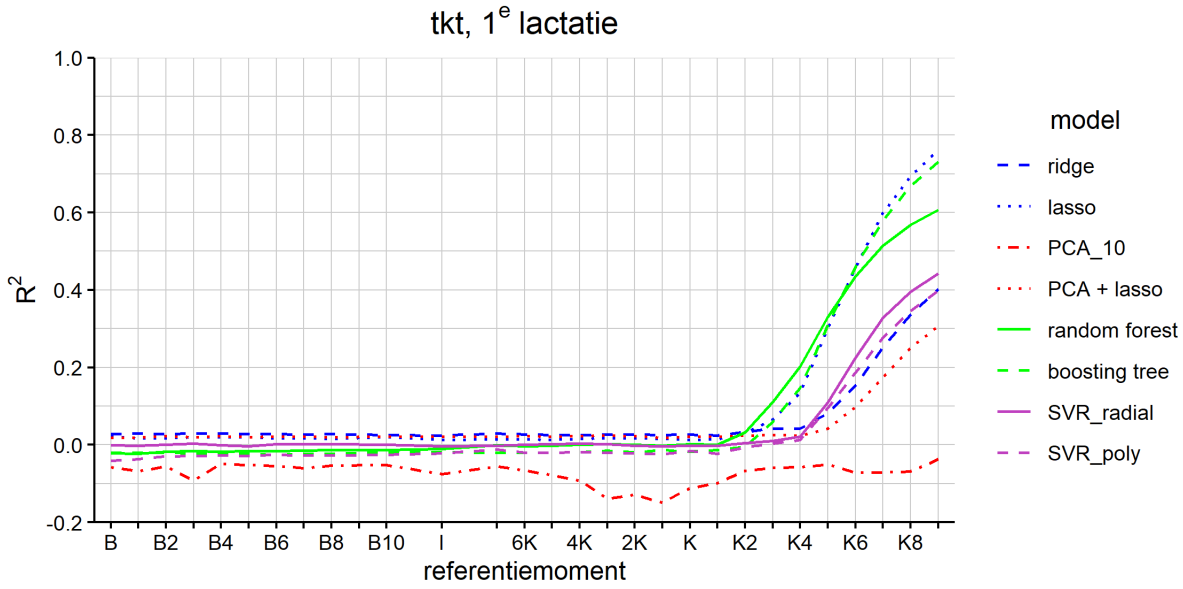


Figuur B.47: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg FPCM

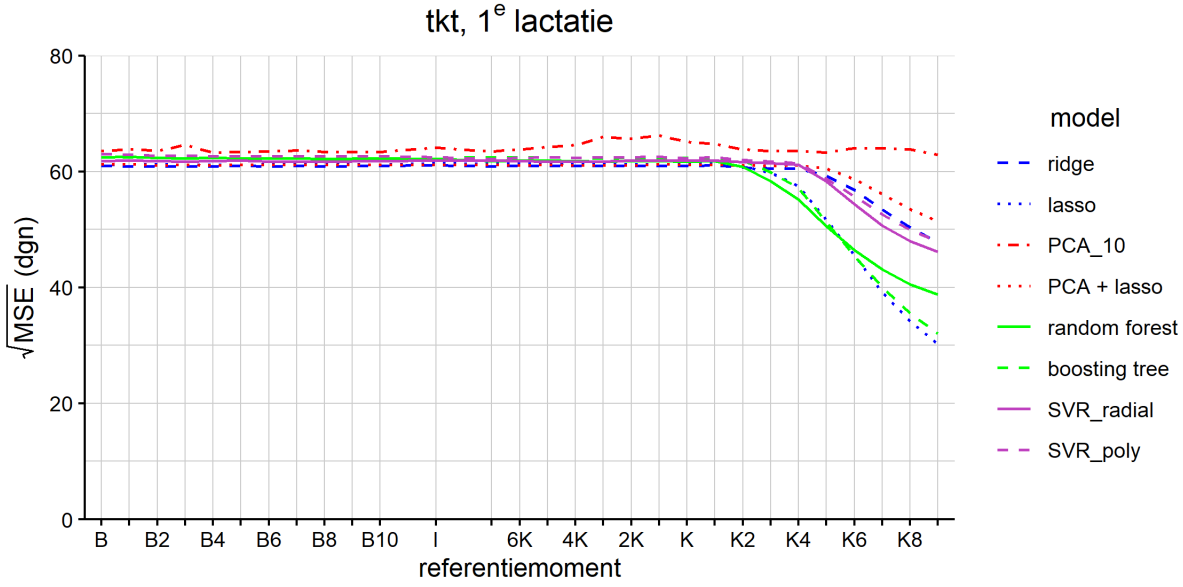


Figuur B.48: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de 305d-productie van de eerste lactatie, uitgedrukt in kg FPCM

- **tkt, 1e lactatie**

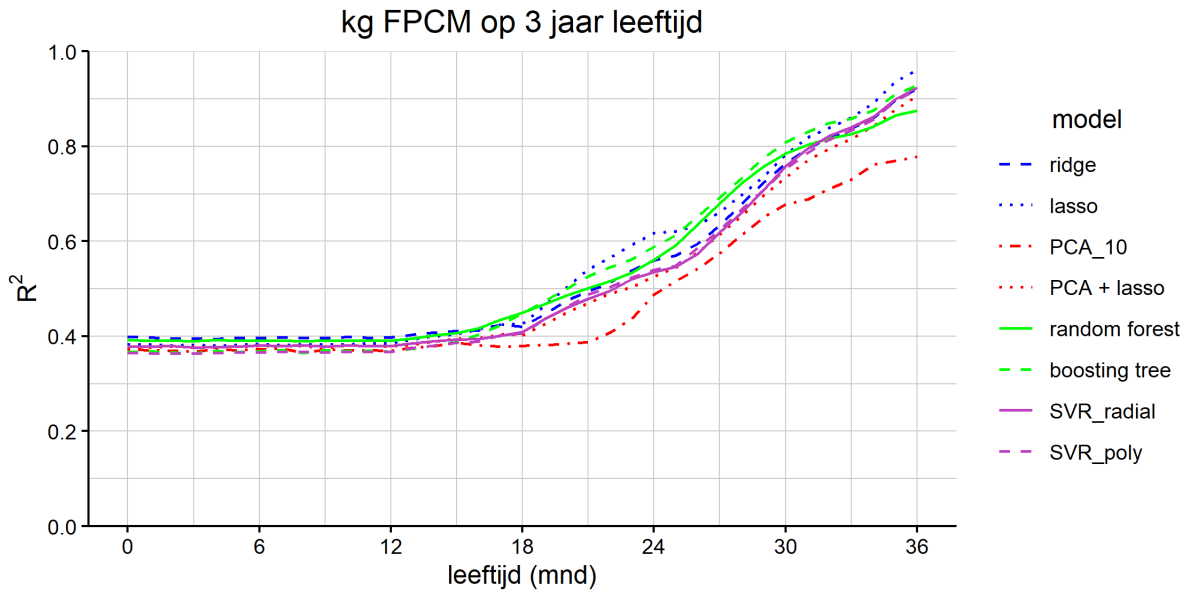


Figuur B.49: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de tussenkalftijd tussen de eerste en de tweede pariteit

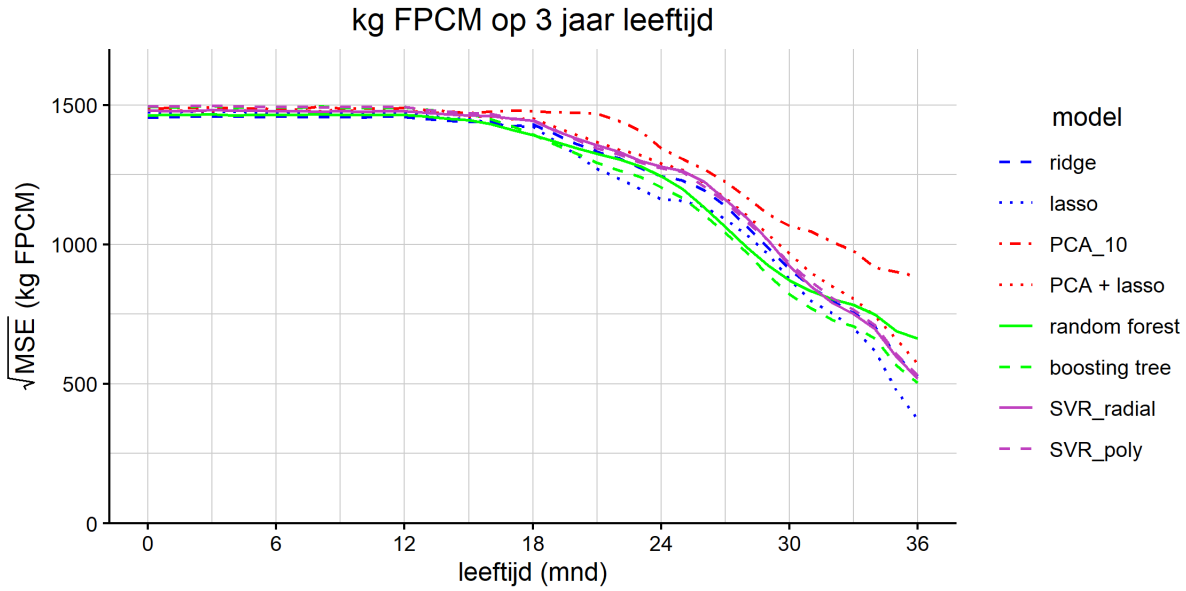


Figuur B.50: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de tussenkalftijd tussen de eerste en de tweede pariteit

- **kg FPCM op 3 jaar leeftijd**

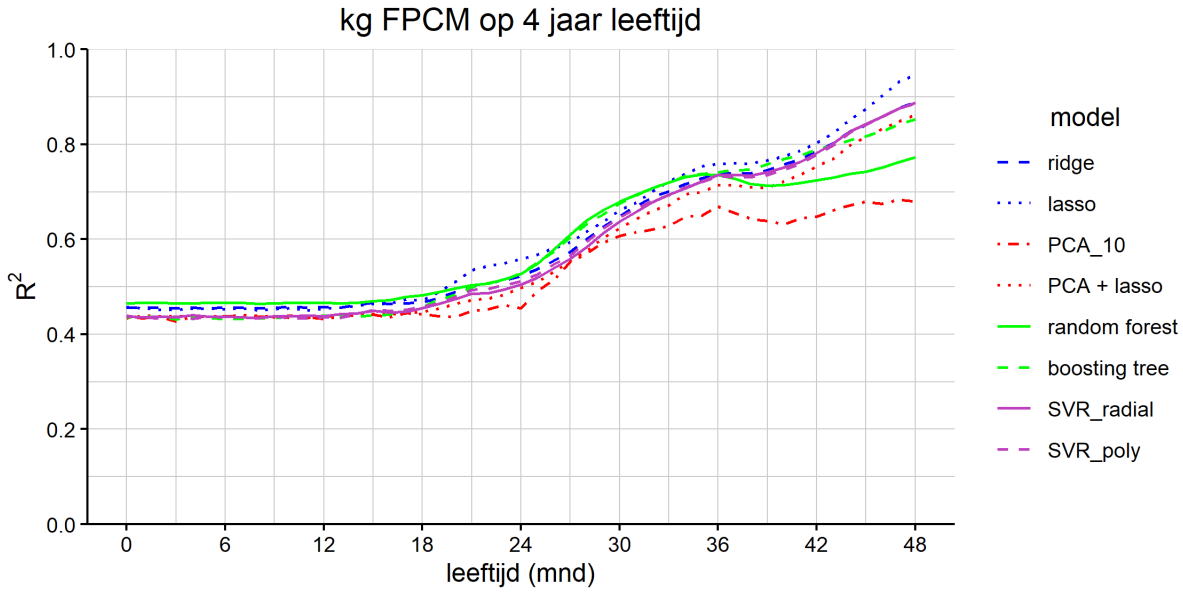


Figuur B.51: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 3 jaar

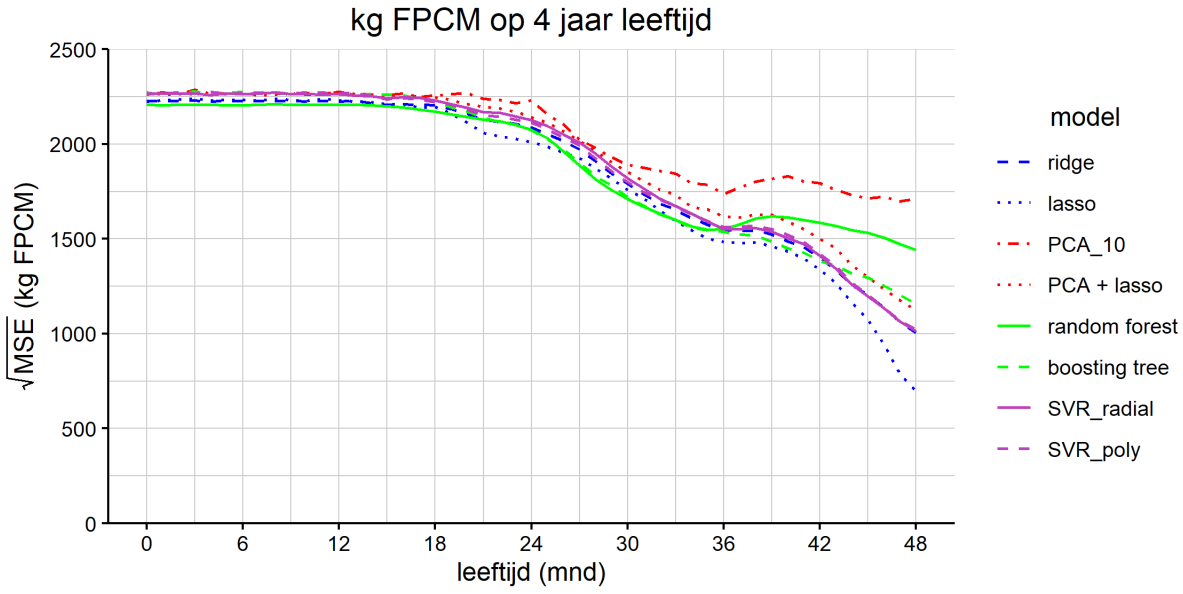


Figuur B.52: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 3 jaar

- **kg FPCM op 4 jaar leeftijd**

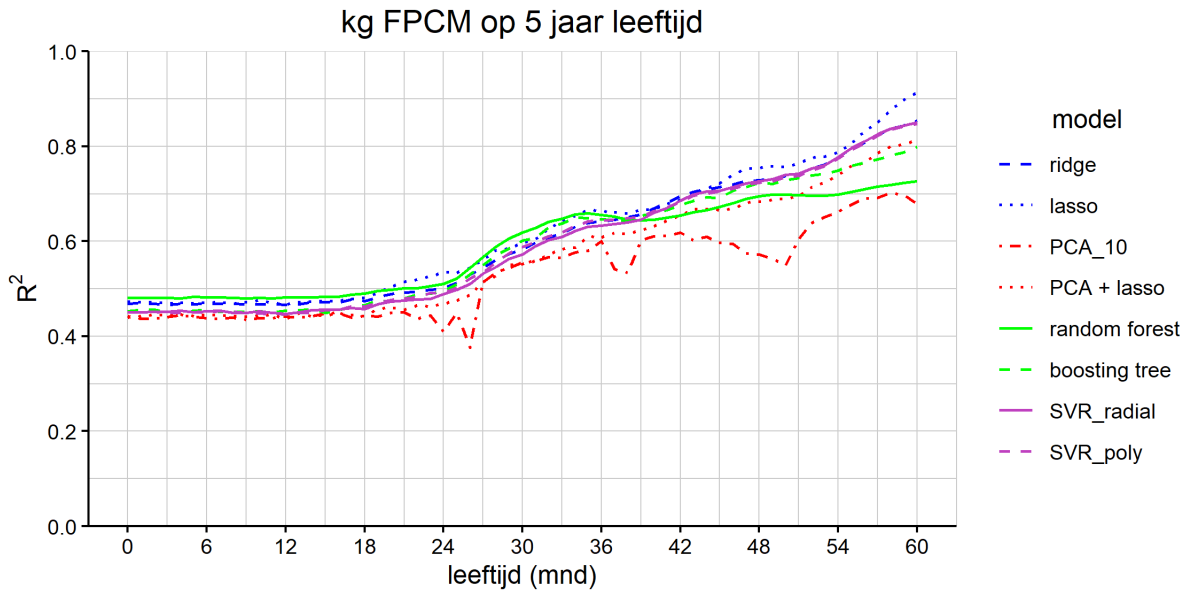


Figuur B.53: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 4 jaar

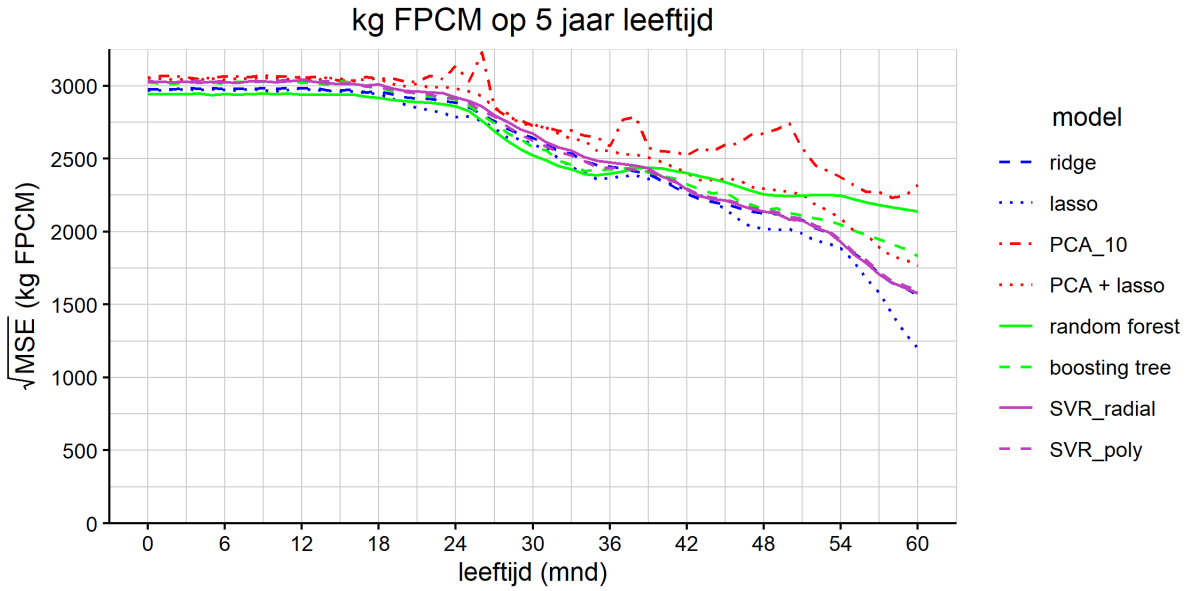


Figuur B.54: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 4 jaar

- **kg FPCM op 5 jaar leeftijd**

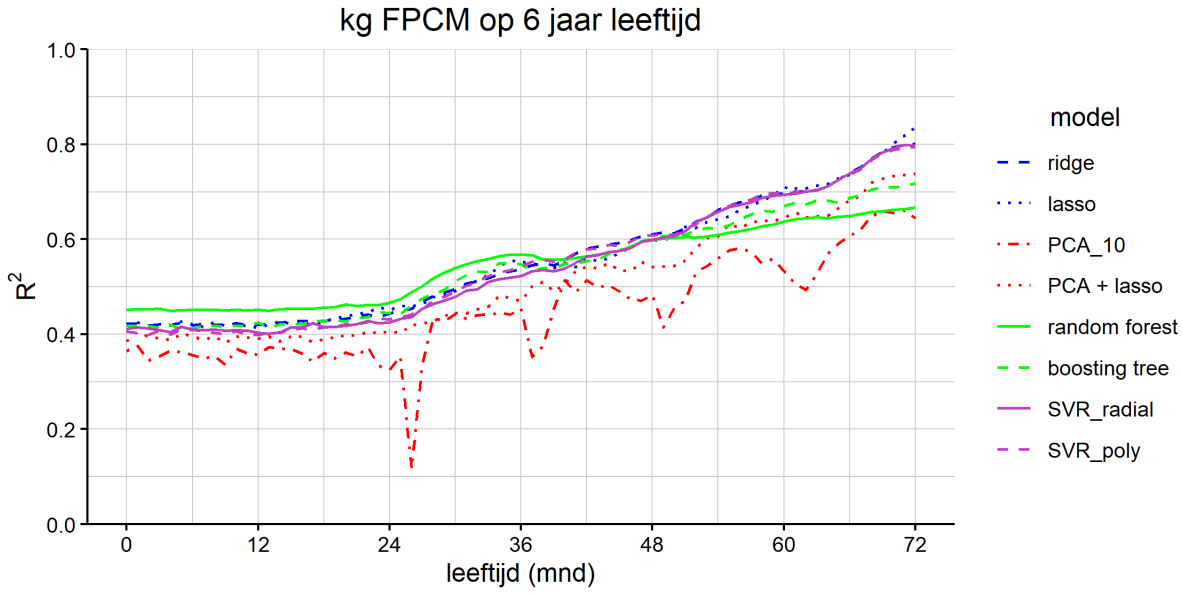


Figuur B.55: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 5 jaar

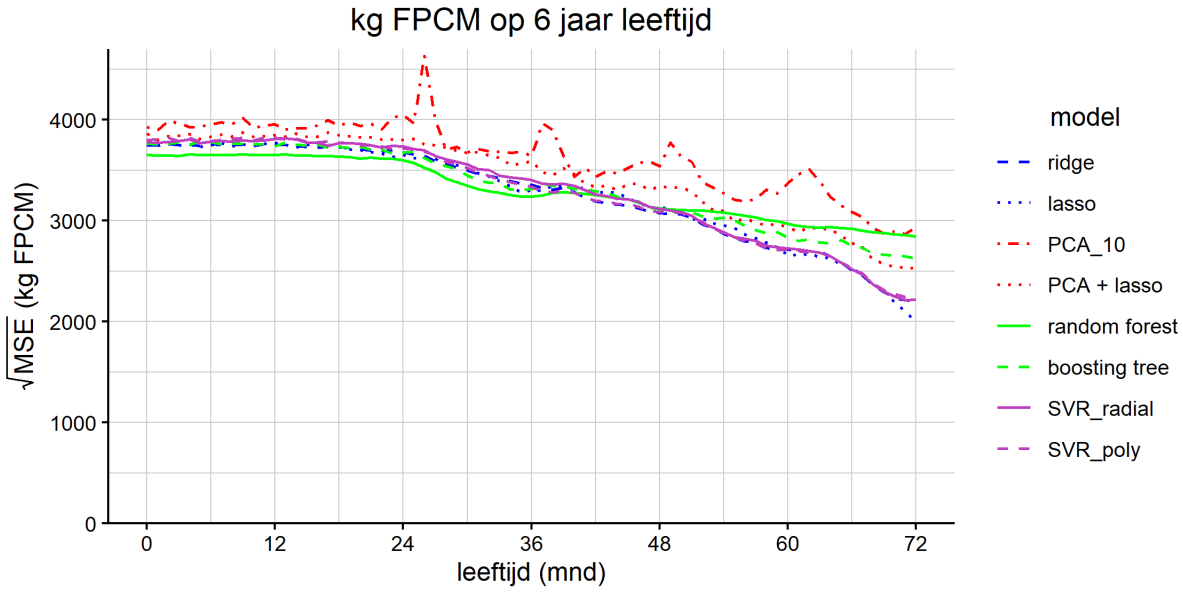


Figuur B.56: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 5 jaar

- **kg FPCM op 6 jaar leeftijd**

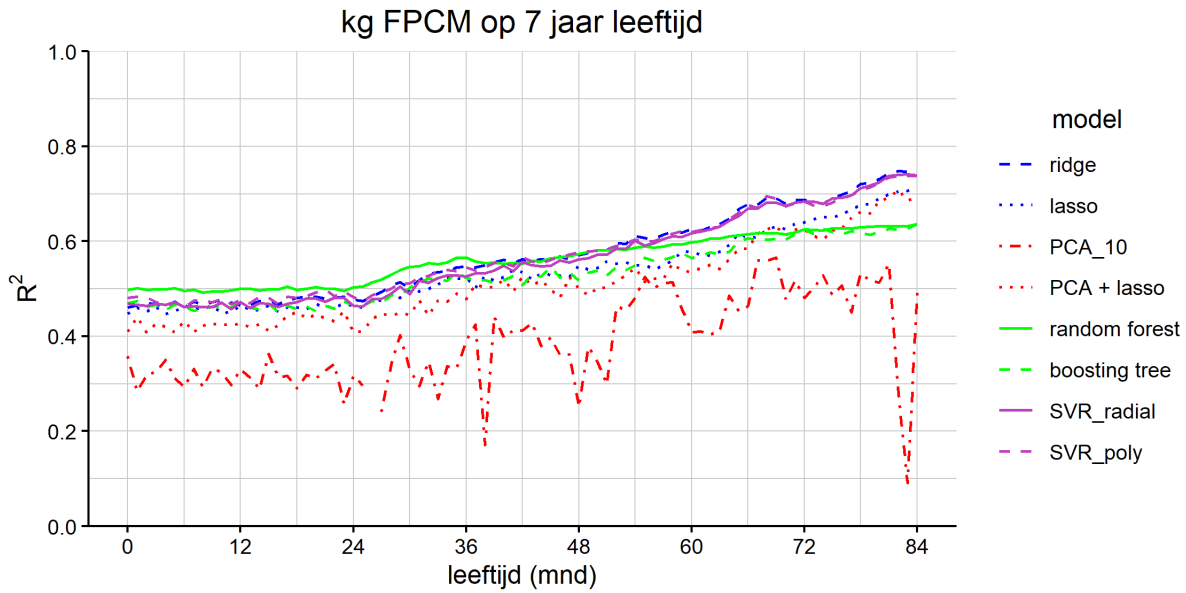


Figuur B.57: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 6 jaar

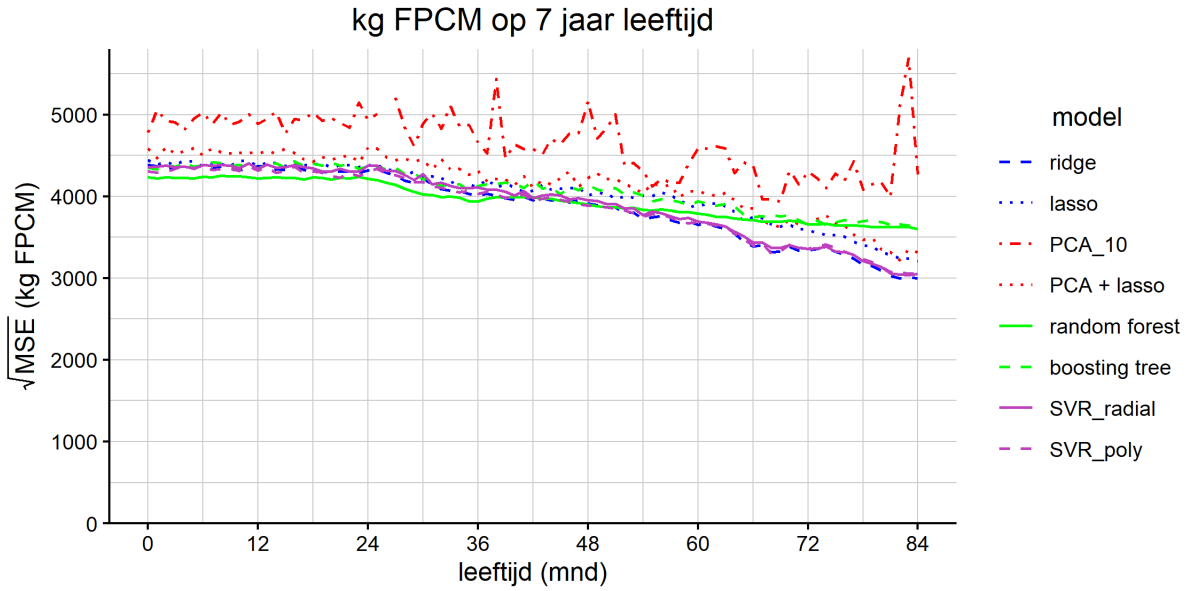


Figuur B.58: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 6 jaar

- **kg FPCM op 7 jaar leeftijd**

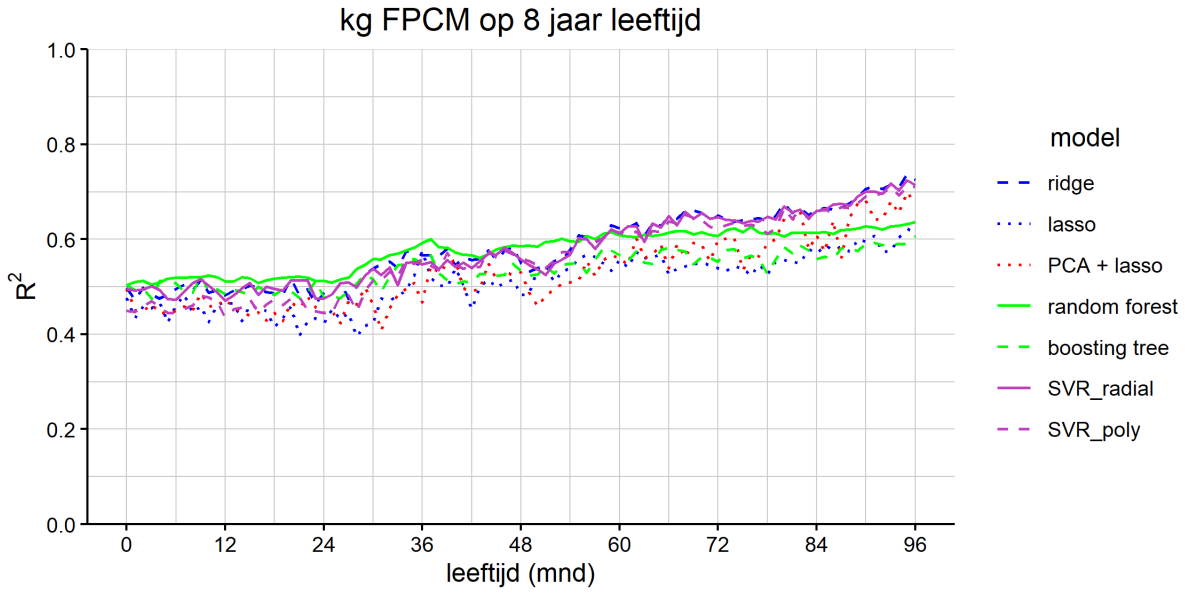


Figuur B.59: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 7 jaar

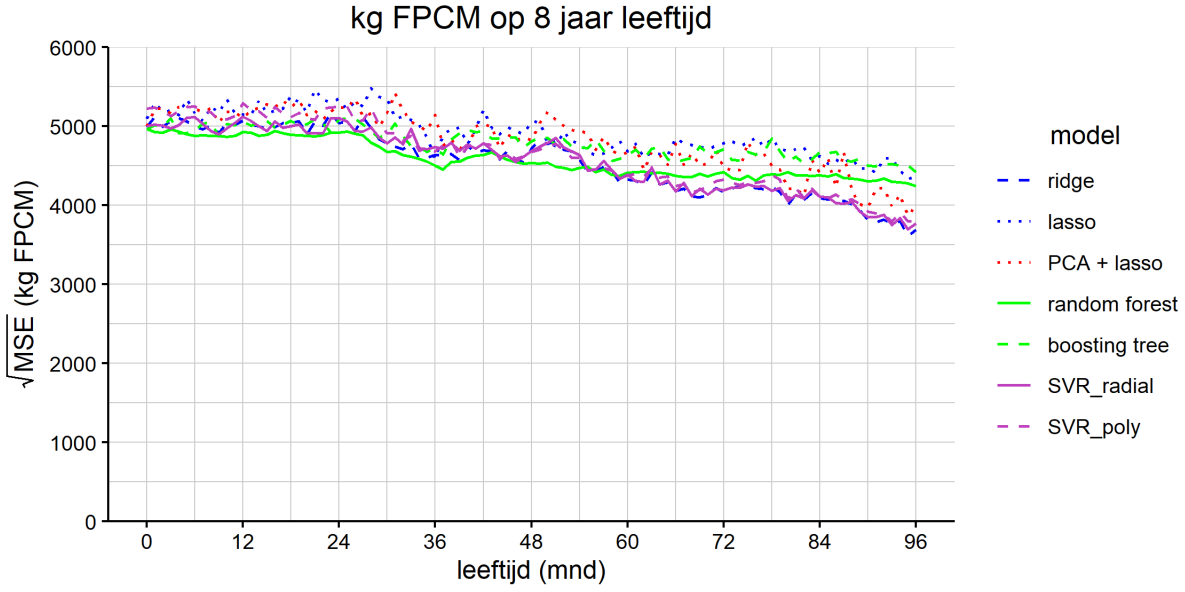


Figuur B.60: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 7 jaar

- **kg FPCM op 8 jaar leeftijd**

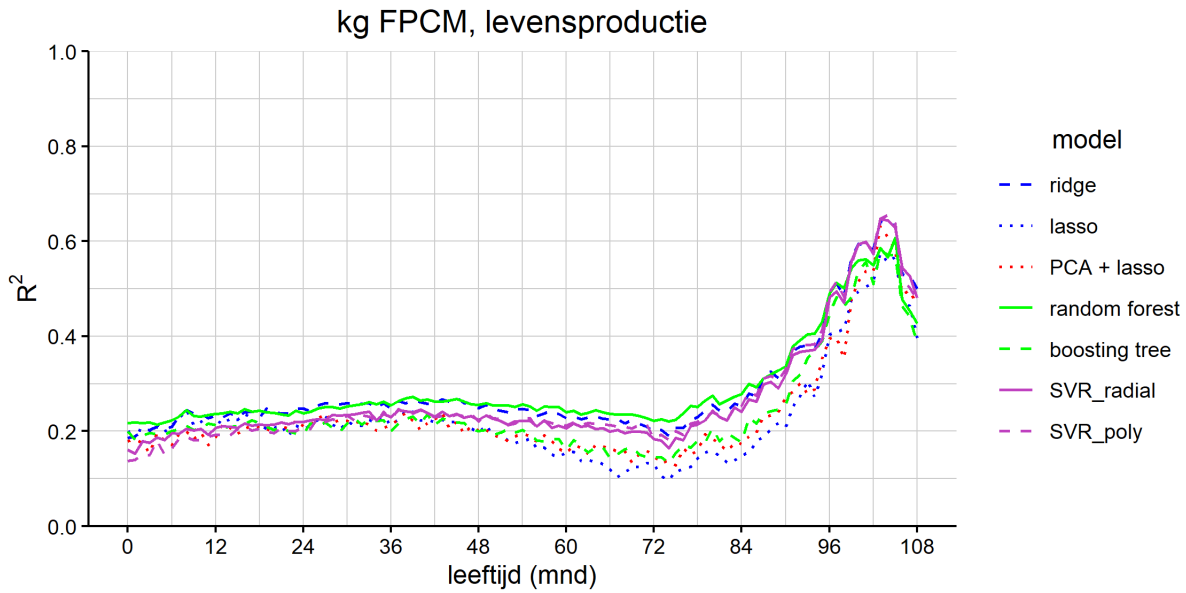


Figuur B.61: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 8 jaar

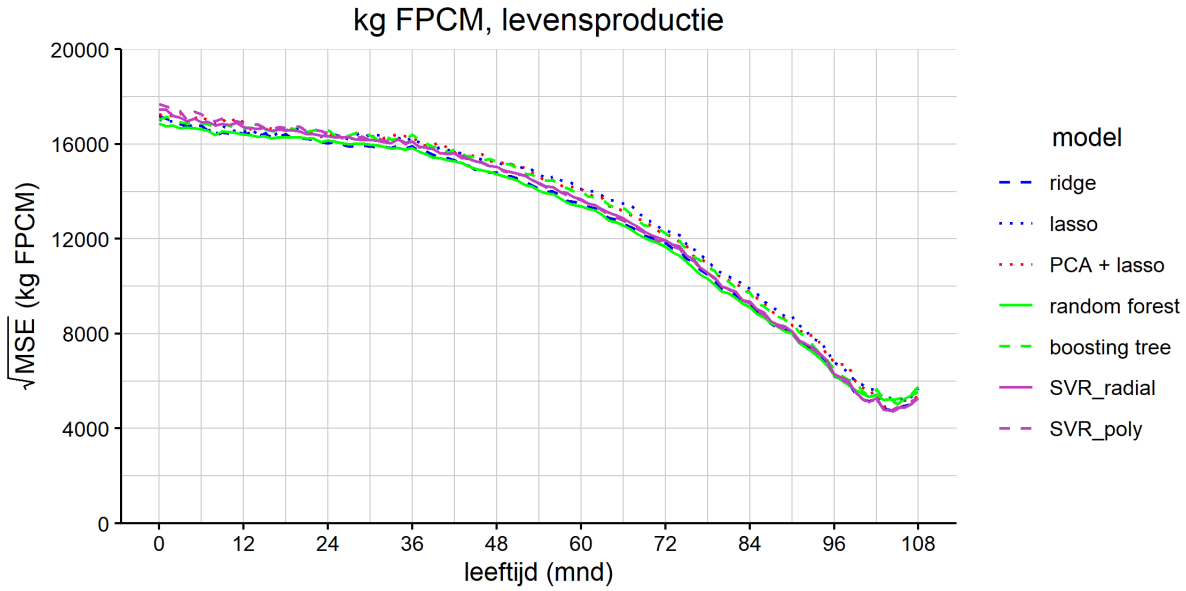


Figuur B.62: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de productie, uitgedrukt in kg FPCM, op een leeftijd van 8 jaar

- **kg FPCM, levensproductie**



Figuur B.63: R^2-waarden op landelijk niveau voor de EVI-modellen met betrekking tot de levensproductie, uitgedrukt in kg FPCM



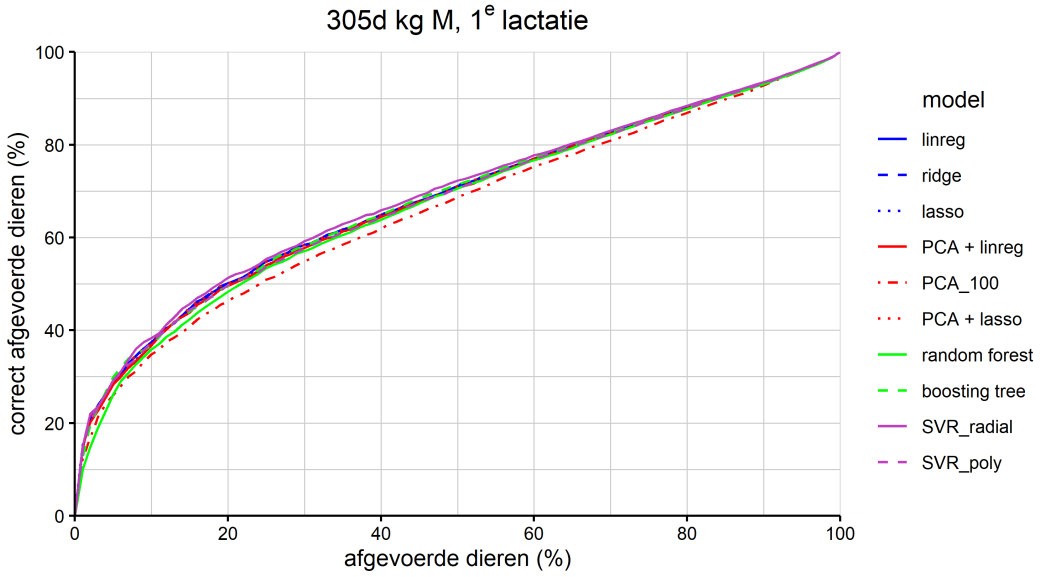
Figuur B.64: Vierkantswortel van de mean squared error (MSE) op landelijk niveau voor de EVI-modellen met betrekking tot de levensproductie, uitgedrukt in kg FPCM

BIJLAGE C

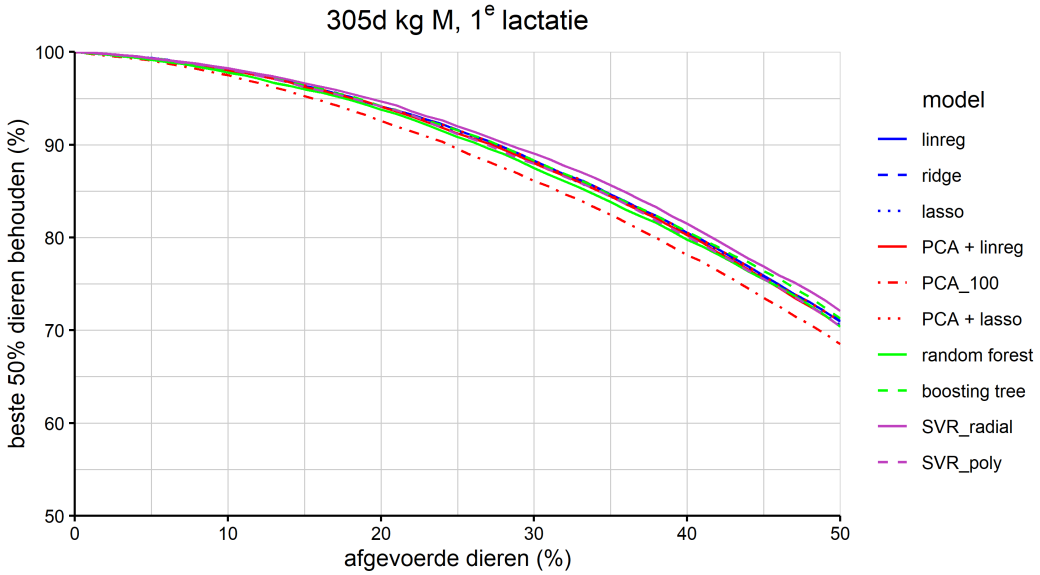
AFVOER VAN OVERTOLLIG JONGVEE: FIGUREN

C.1 EVA: één model op landelijk niveau

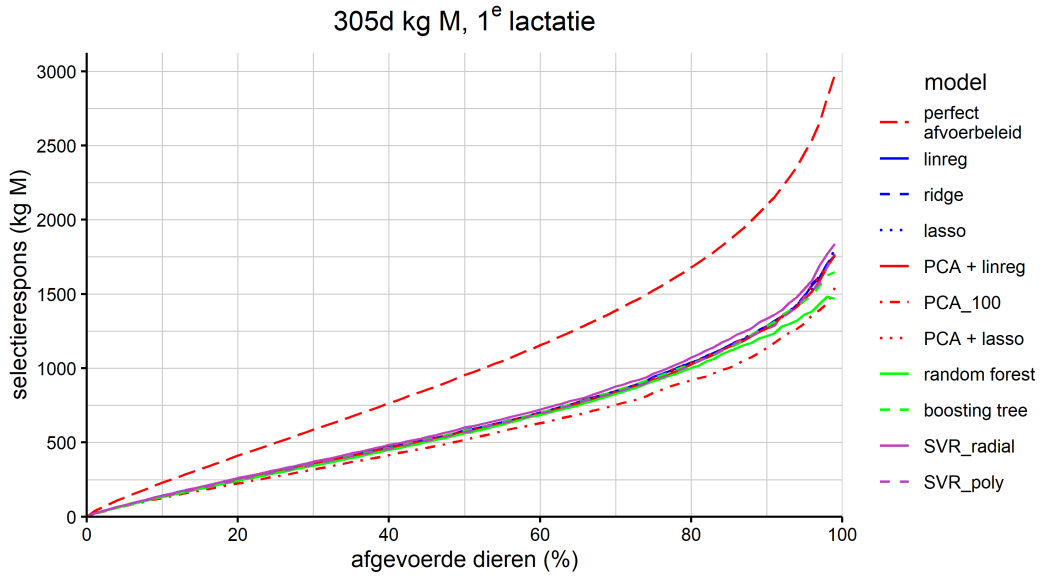
- **305d kg M, 1e lactatie**



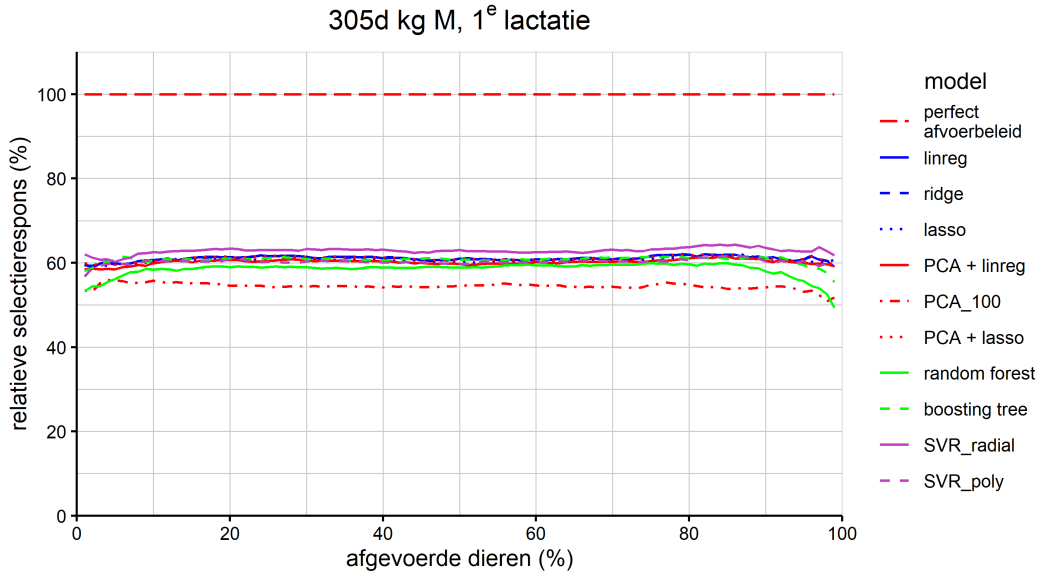
Figuur C.1: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg melk (EVA-modellen)



Figuur C.2: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg melk (EVA-modellen)

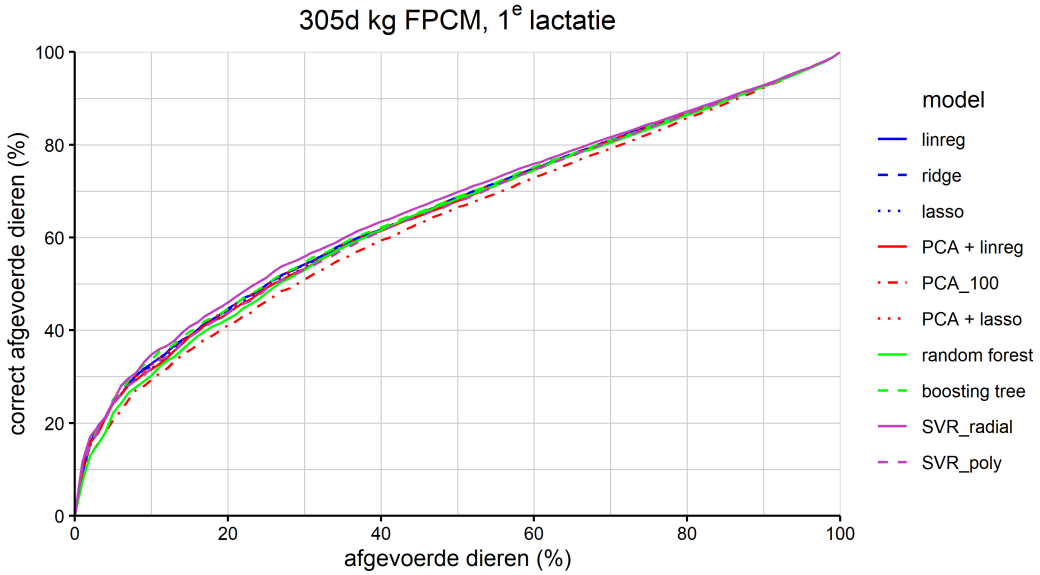


Figuur C.3: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg melk (EVA-modellen)

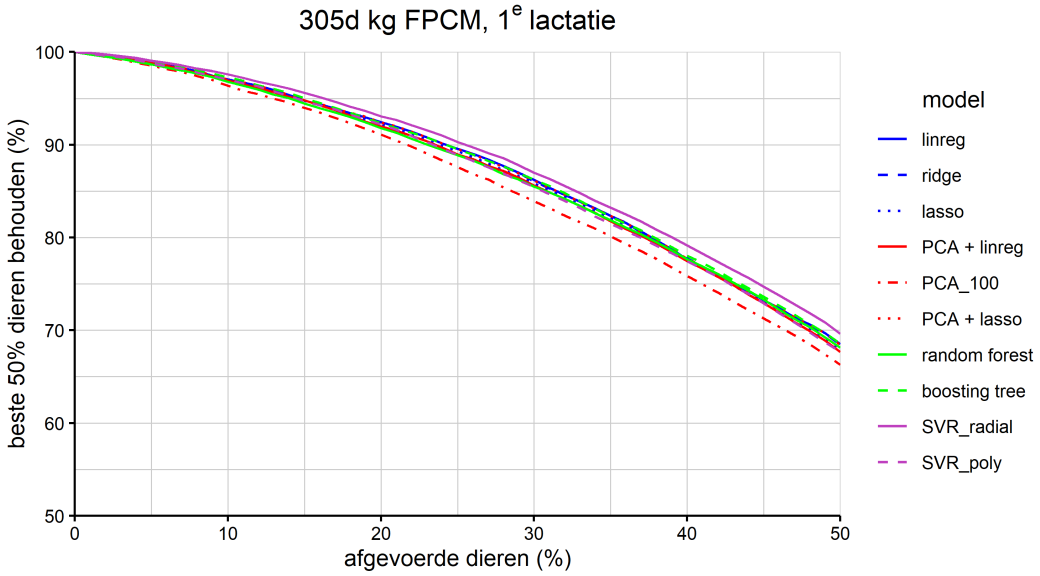


Figuur C.4: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg melk (EVA-modellen)

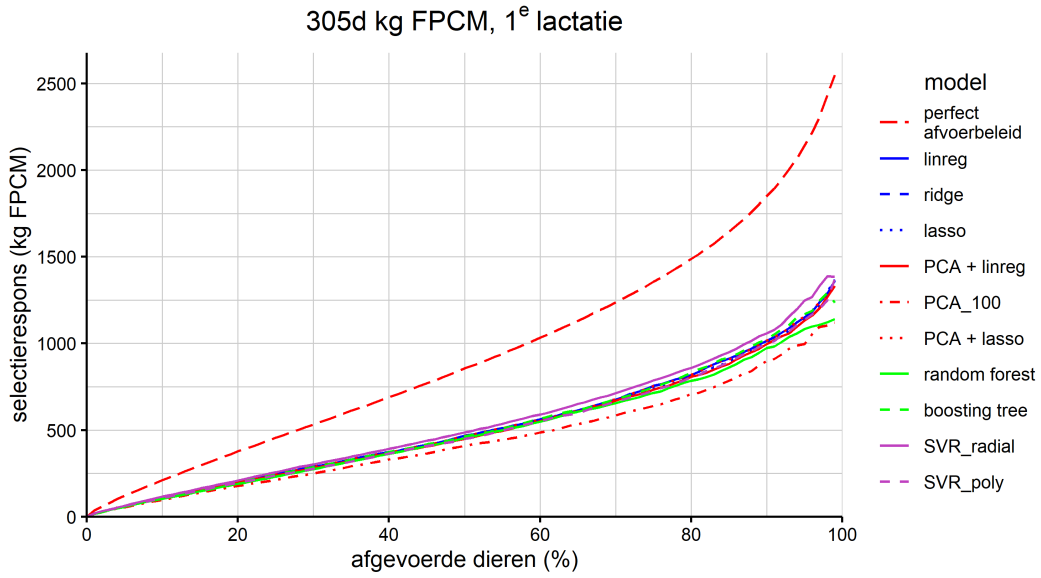
- **305d kg FPCM, 1e lactatie**



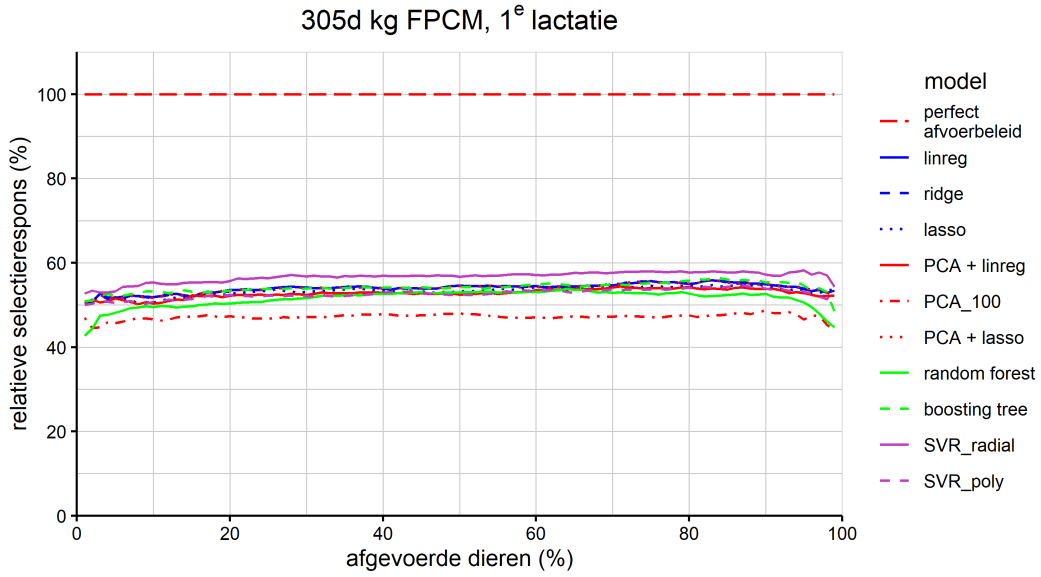
Figuur C.5: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg FPCM (EVA-modellen)



Figuur C.6: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg FPCM (EVA-modellen)

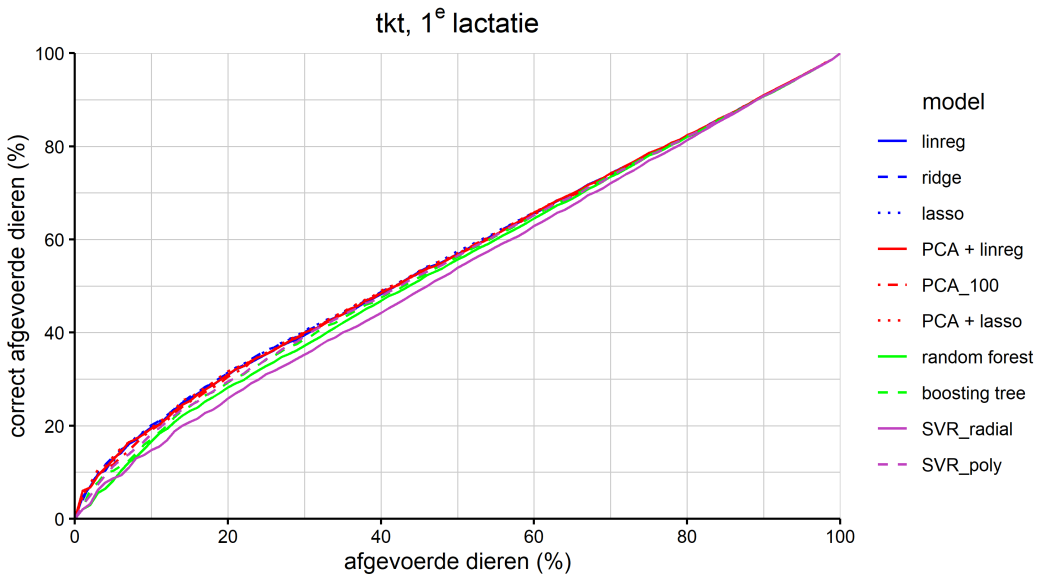


Figuur C.7: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg FPCM (EVA-modellen)

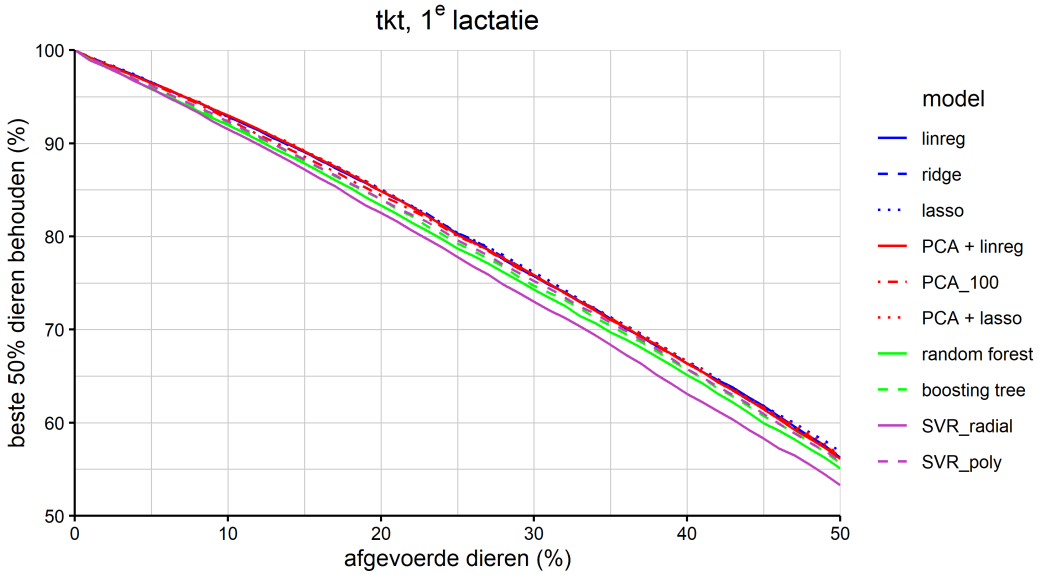


Figuur C.8: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg FPCM (EVA-modellen)

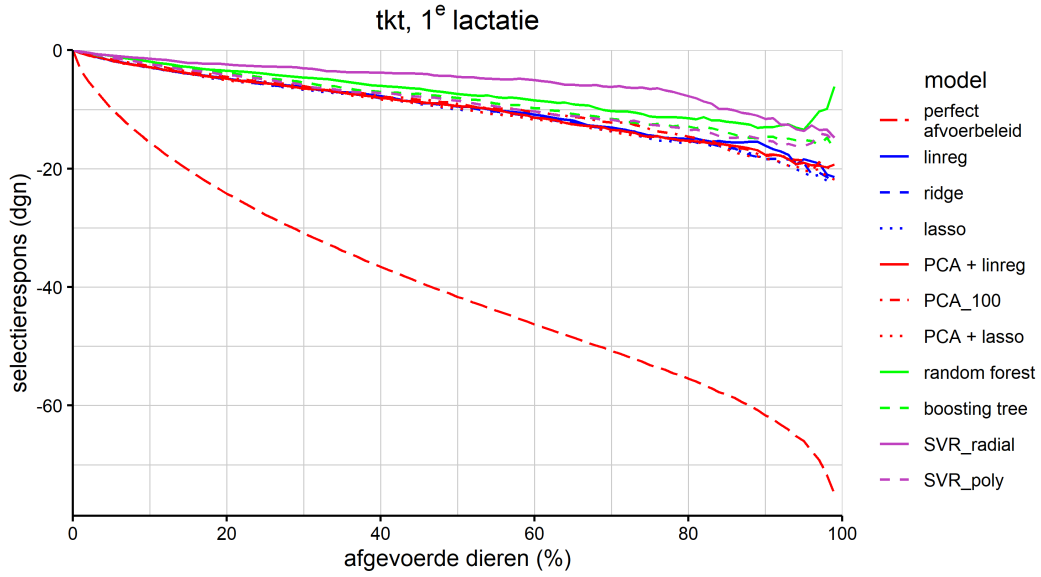
- **tkt, 1e lactatie**



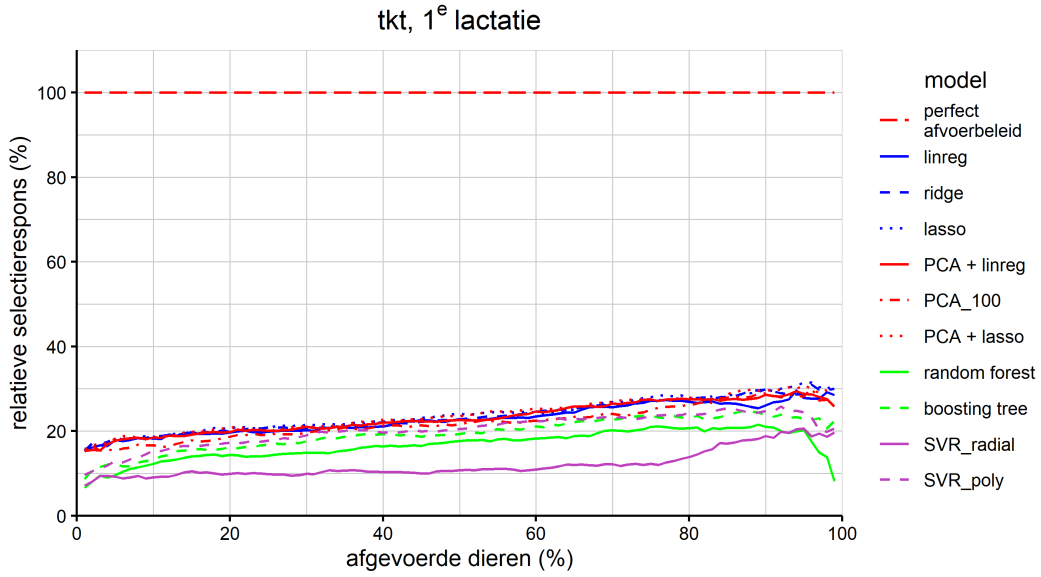
Figuur C.9: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde tussenkalftijden tussen de eerste en de tweede pariteit (EVA-modellen)



Figuur C.10: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde tussenkalftijden tussen de eerste en de tweede pariteit (EVA-modellen)

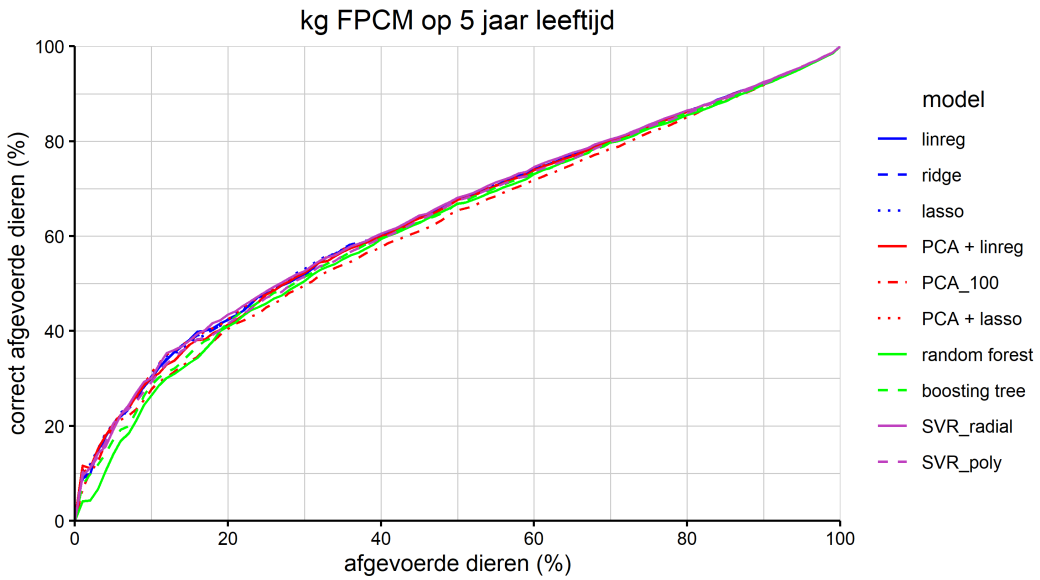


Figuur C.11: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde tussenkalftijden tussen de eerste en de tweede pariteit (EVA-modellen)

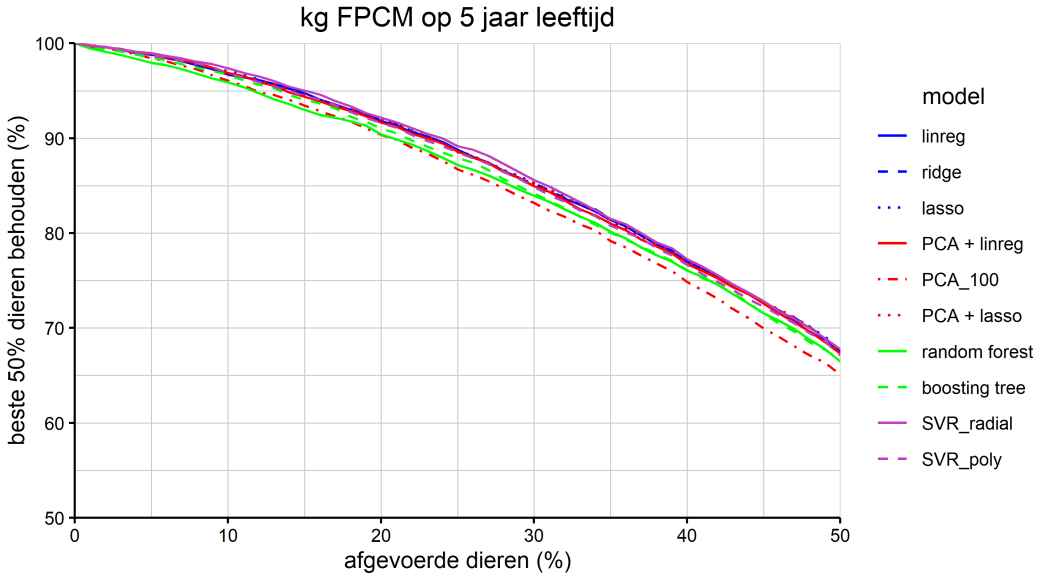


Figuur C.12: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde tussenkalftijden tussen de eerste en de tweede pariteit (EVA-modellen)

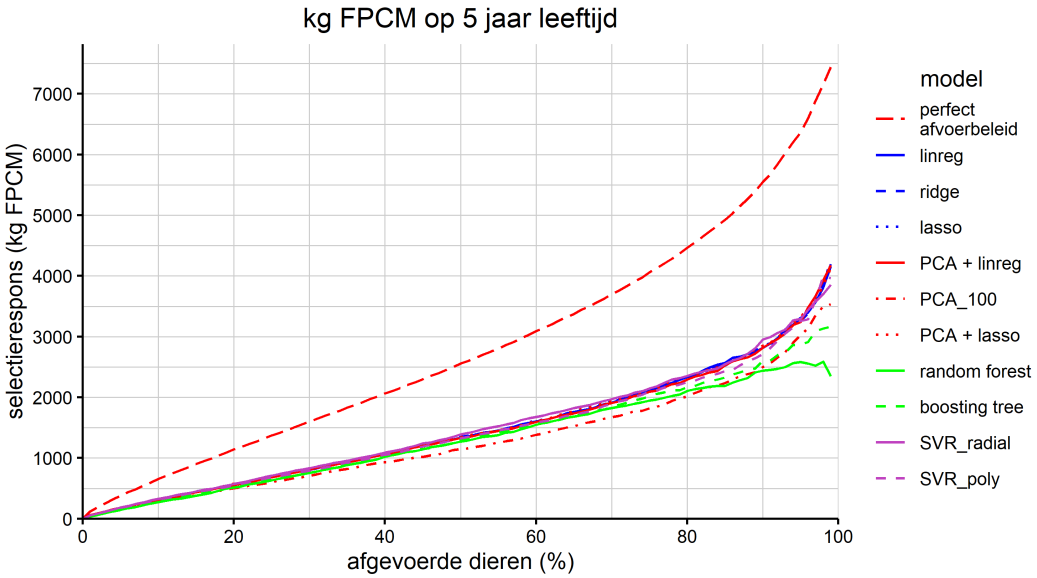
- **kg FPCM op 5 jaar leeftijd**



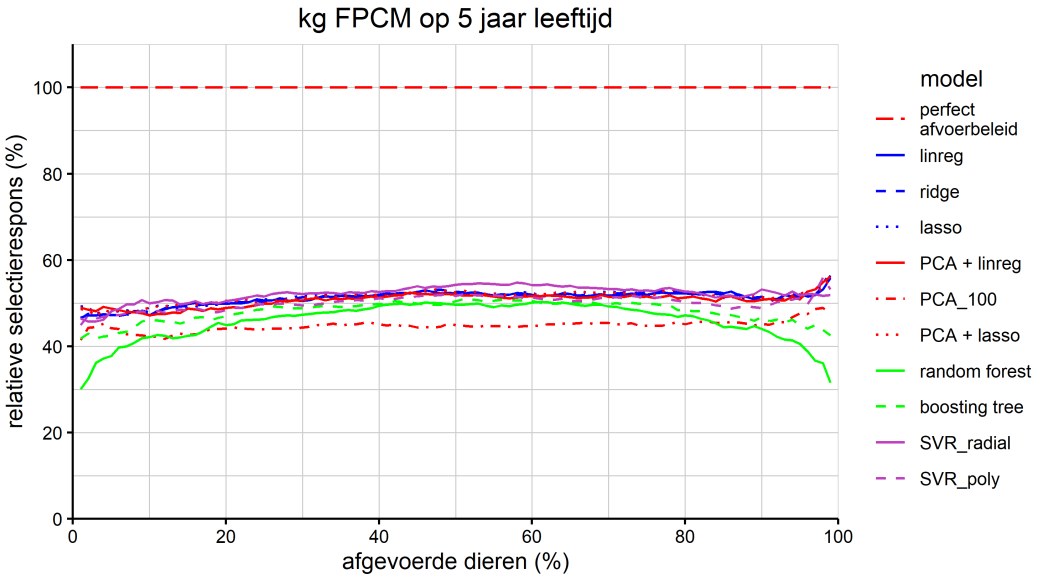
Figuur C.13: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde kg FPCM-producties op een leeftijd van 5 jaar (EVA-modellen)



Figuur C.14: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde kg FPCM-producties op een leeftijd van 5 jaar (EVA-modellen)

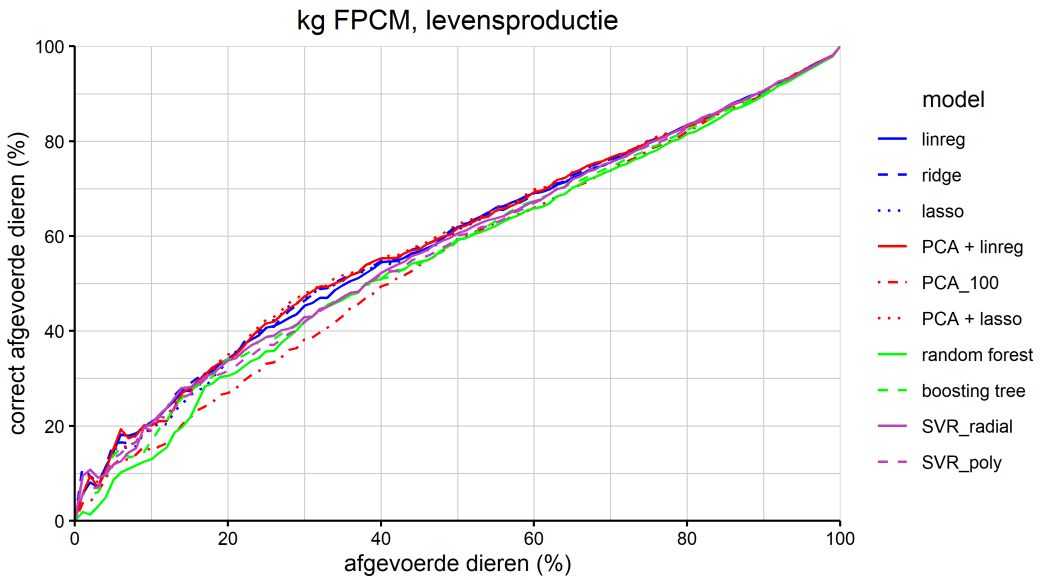


Figuur C.15: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde kg FPCM-producties op een leeftijd van 5 jaar (EVA-modellen)

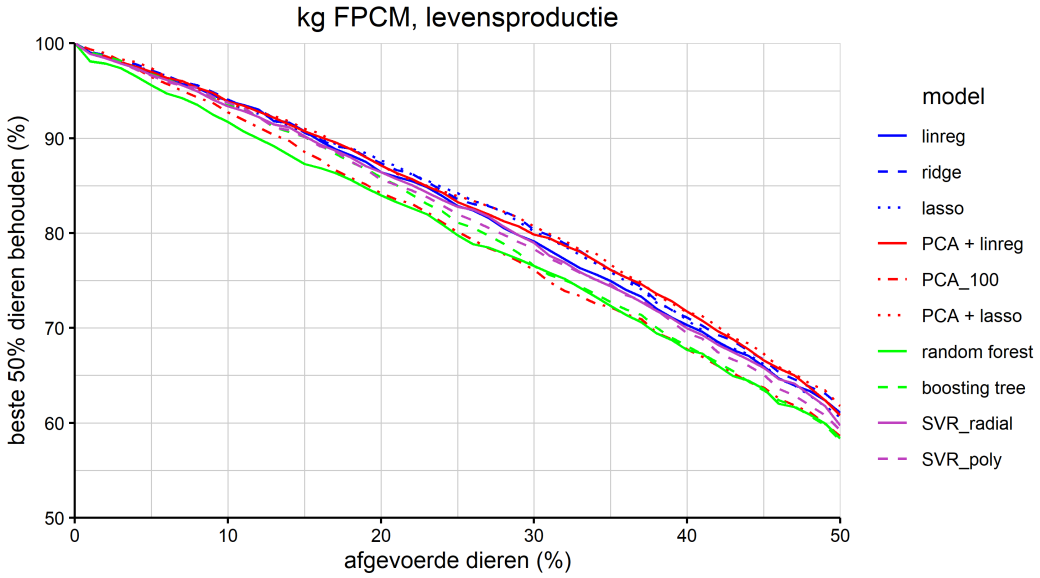


Figuur C.16: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde kg FPCM-producties op een leeftijd van 5 jaar (EVA-modellen)

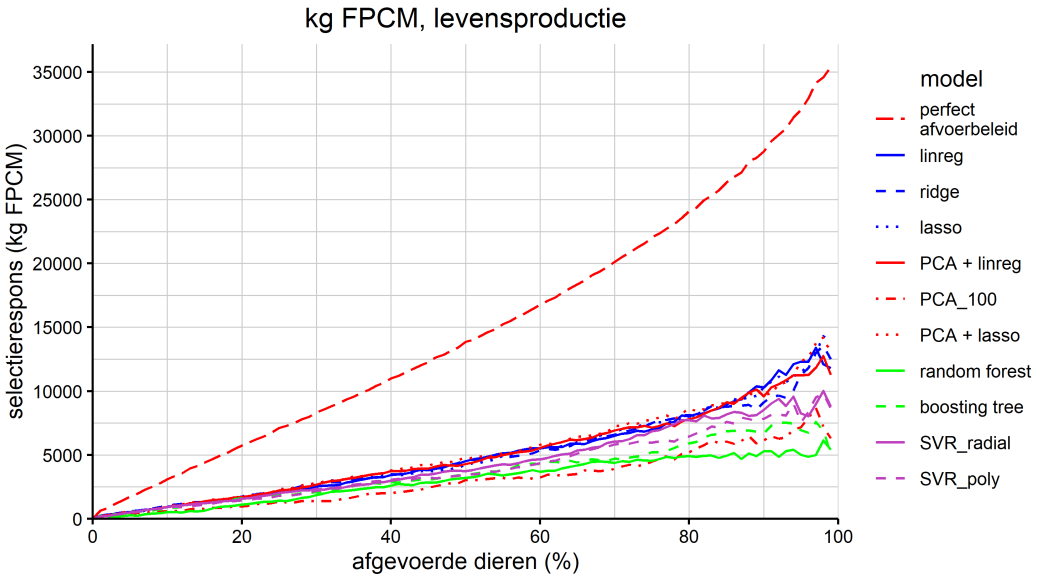
- **kg FPCM,levensproductie**



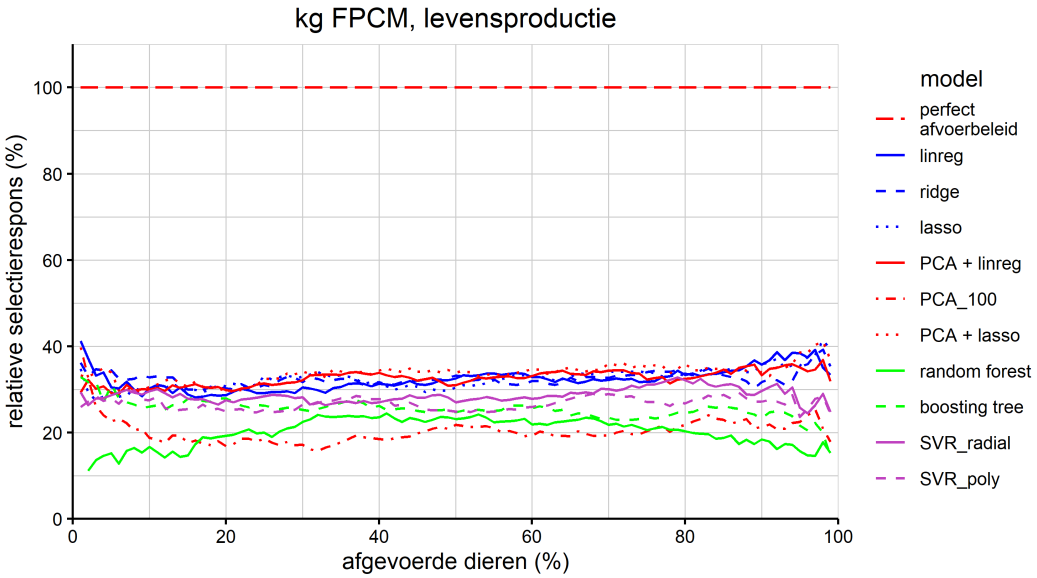
Figuur C.17: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM (EVA-modellen)



Figuur C.18: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM (EVA-modellen)



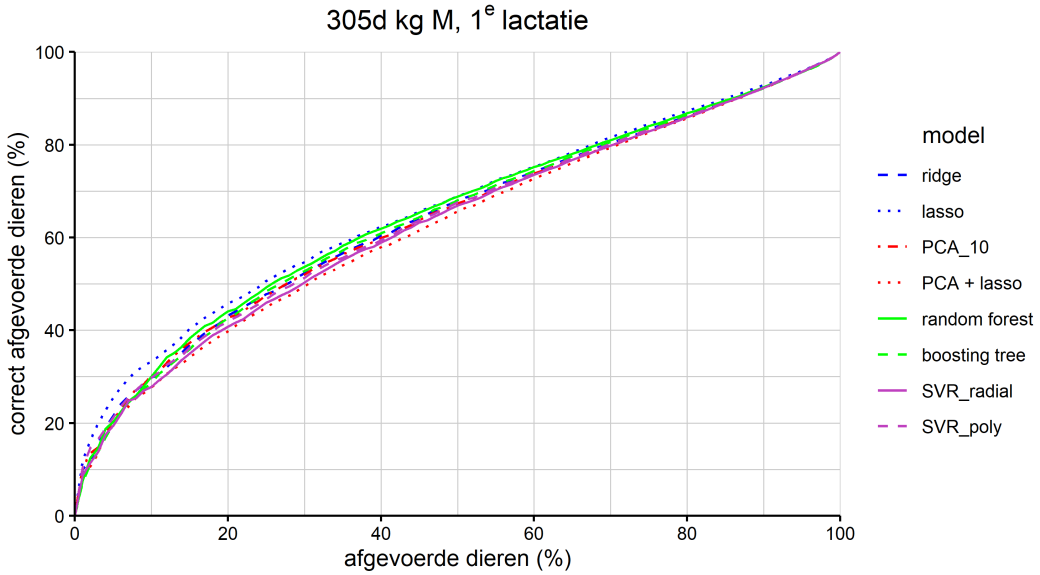
Figuur C.19: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM (EVA-modellen)



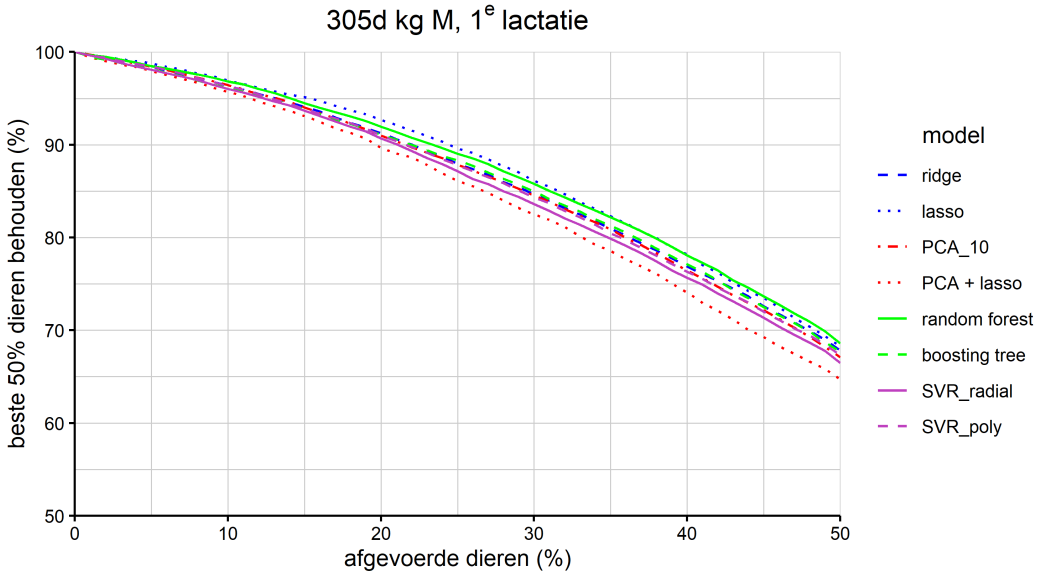
Figuur C.20: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM (EVA-modellen)

C.2 EVI: één model voor ieder bedrijf

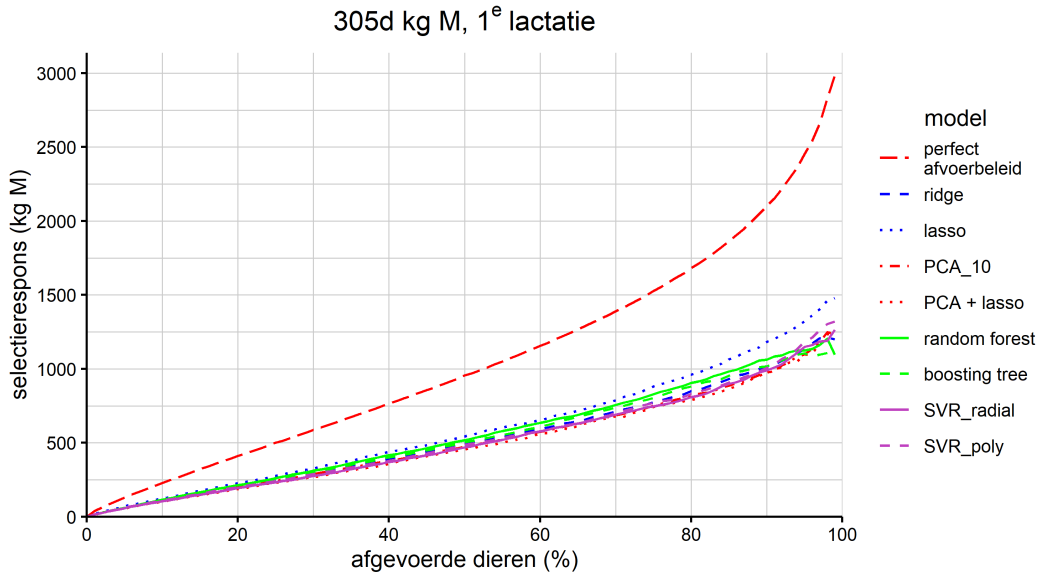
- **305d kg M, 1e lactatie**



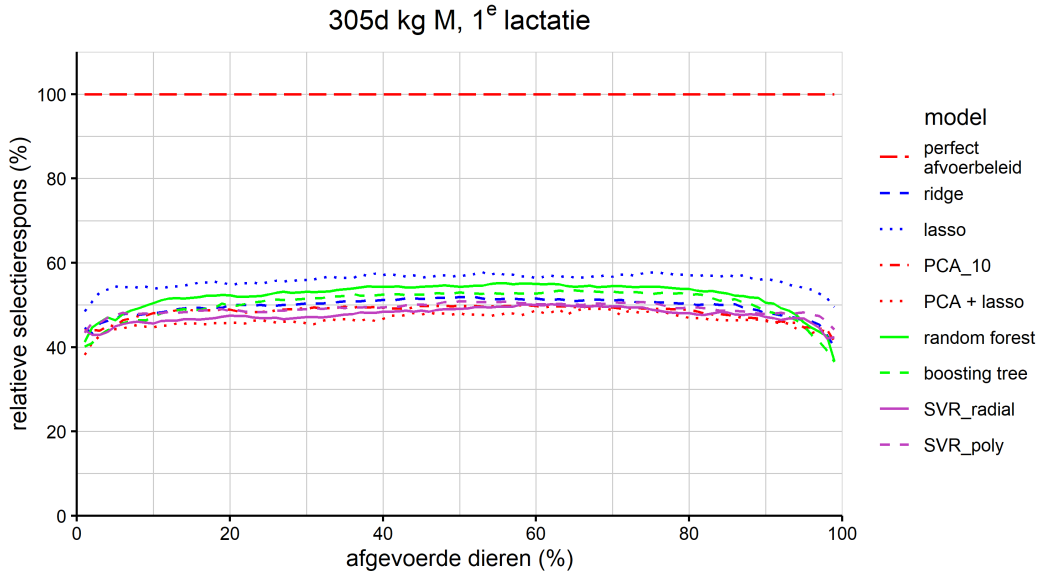
Figuur C.21: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg melk (EVI-modellen)



Figuur C.22: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg melk (EVI-modellen)

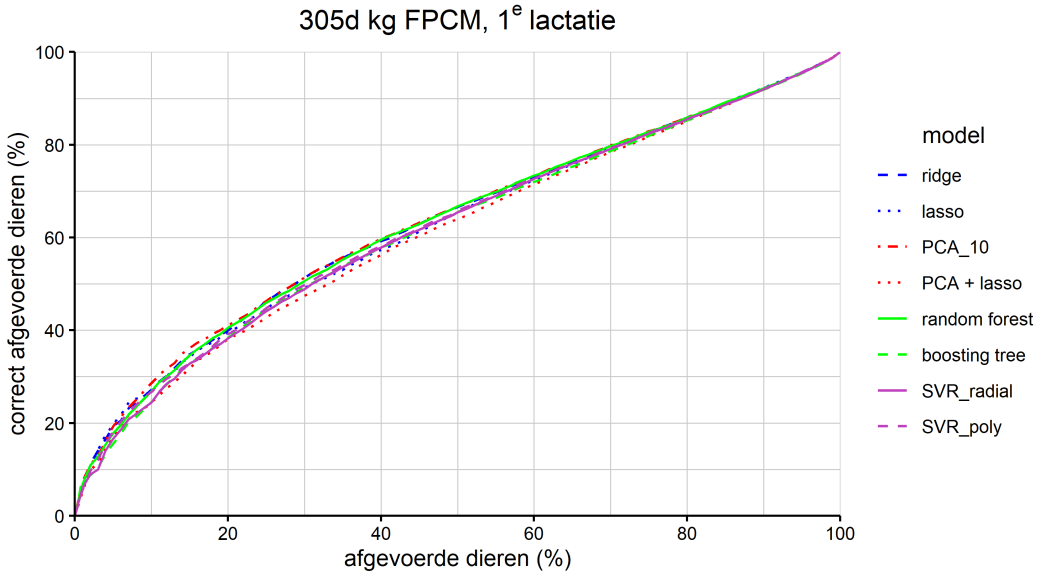


Figuur C.23: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg melk (EVI-modellen)

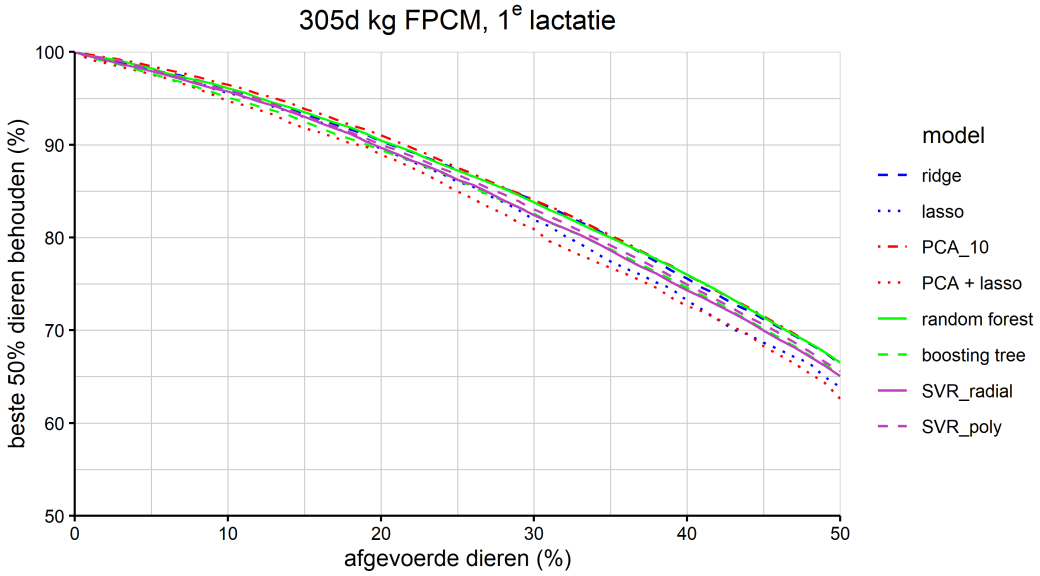


Figuur C.24: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg melk (EVI-modellen)

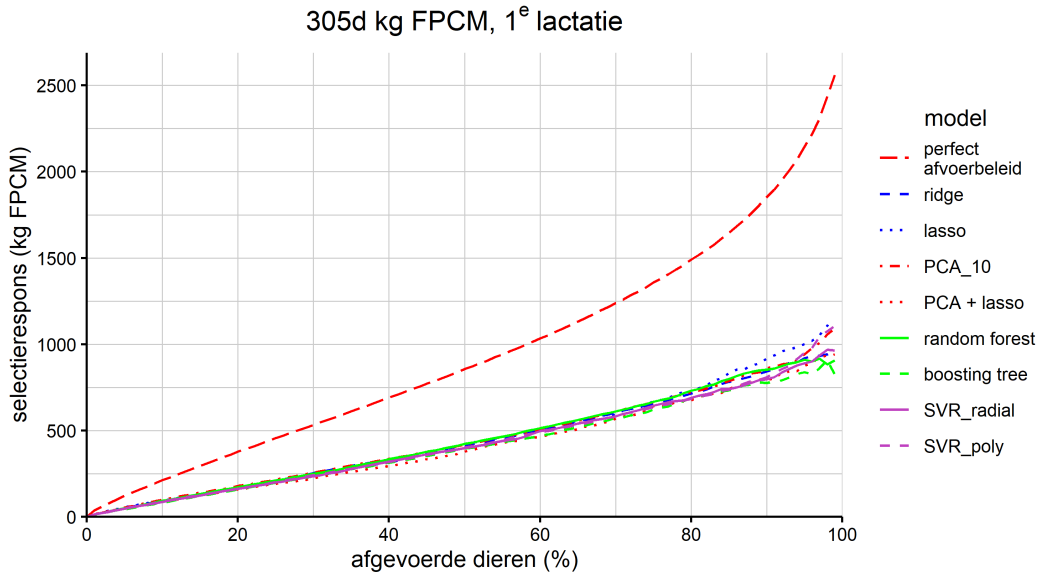
- **305d kg FPCM, 1e lactatie**



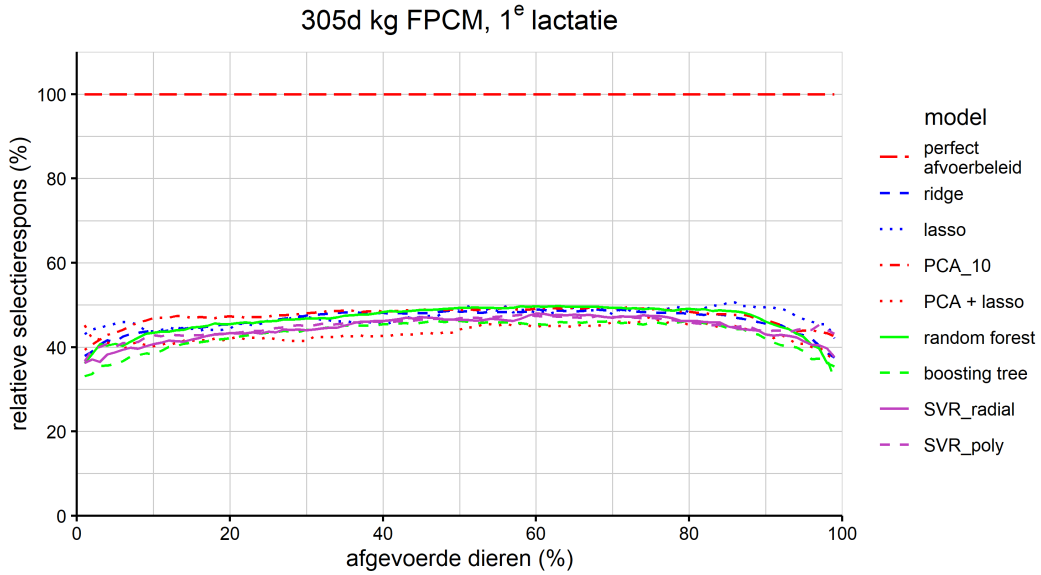
Figuur C.25: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg FPCM (EVI-modellen)



Figuur C.26: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg FPCM (EVI-modellen)

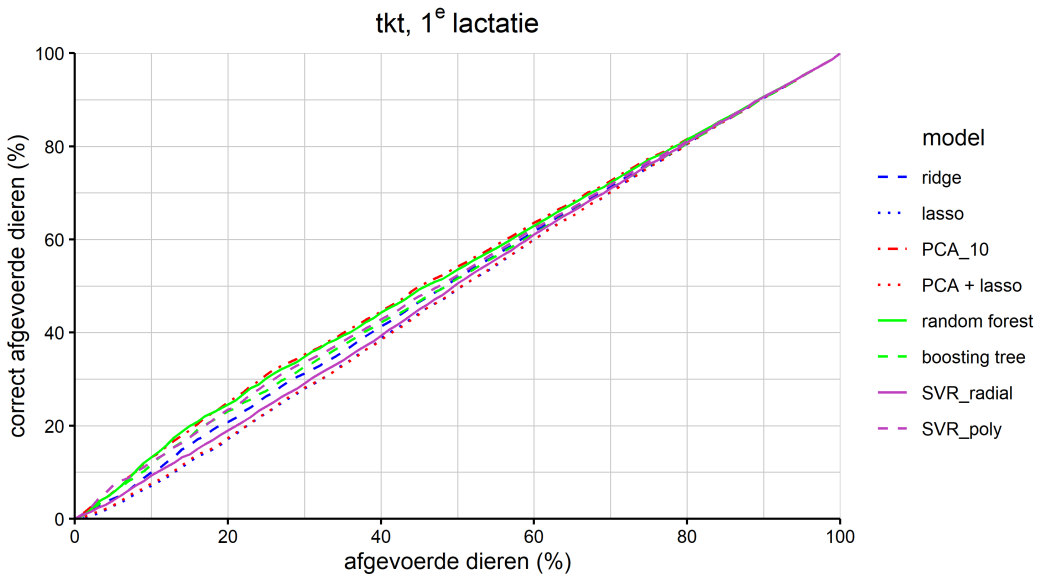


Figuur C.27: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg FPCM (EVI-modellen)

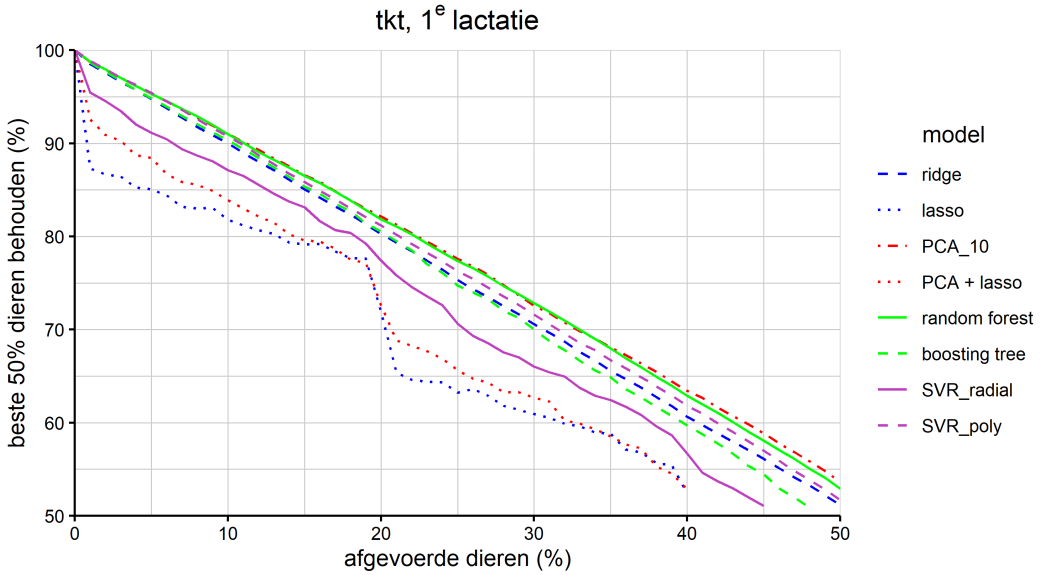


Figuur C.28: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg FPCM (EVI-modellen)

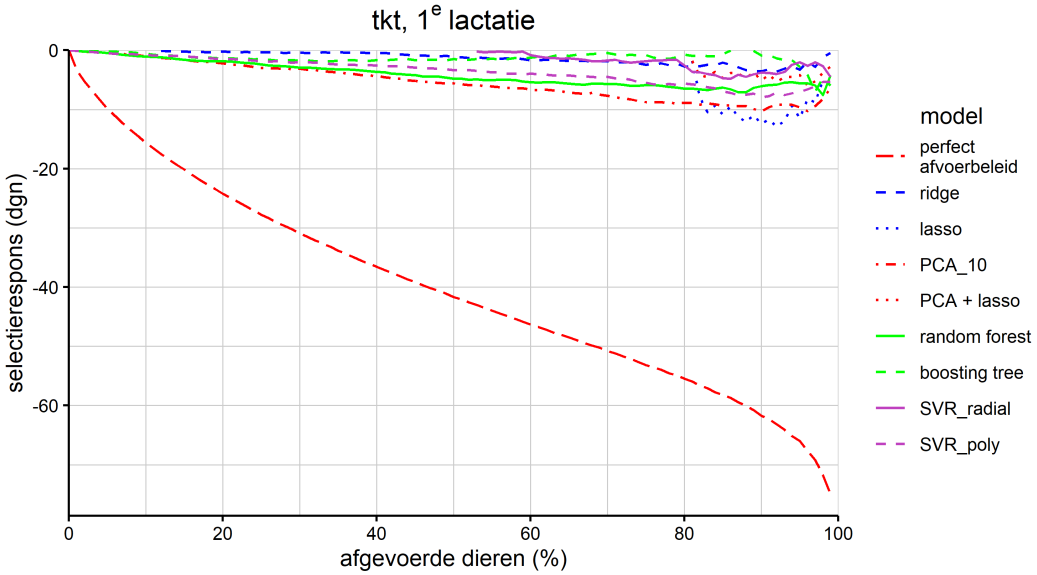
- **tkt, 1e lactatie**



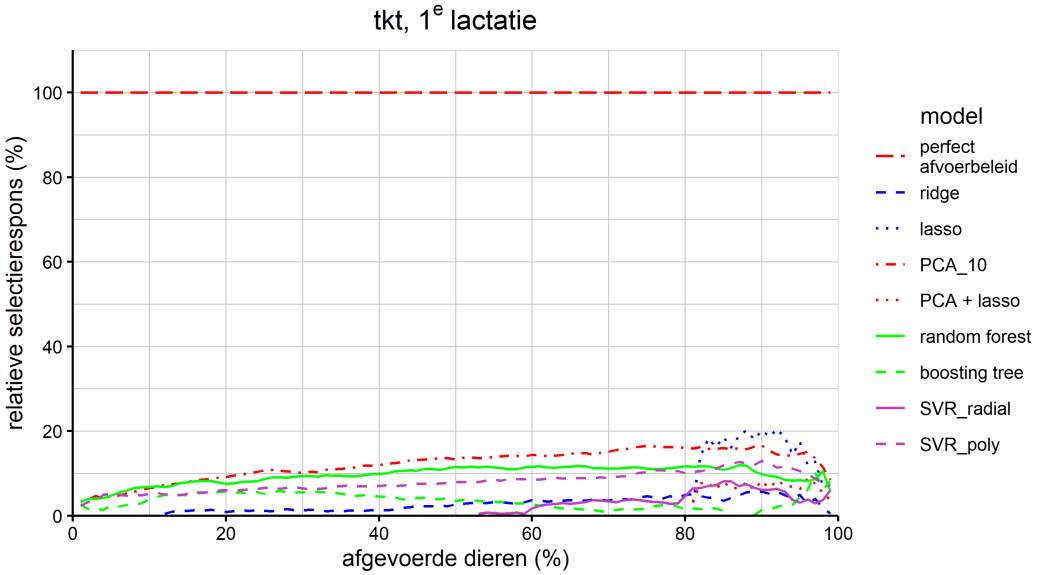
Figuur C.29: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde tussenkalftijden tussen de eerste en de tweede pariteit (EVI-modellen)



Figuur C.30: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde tussenkalftijden tussen de eerste en de tweede pariteit (EVI-modellen)

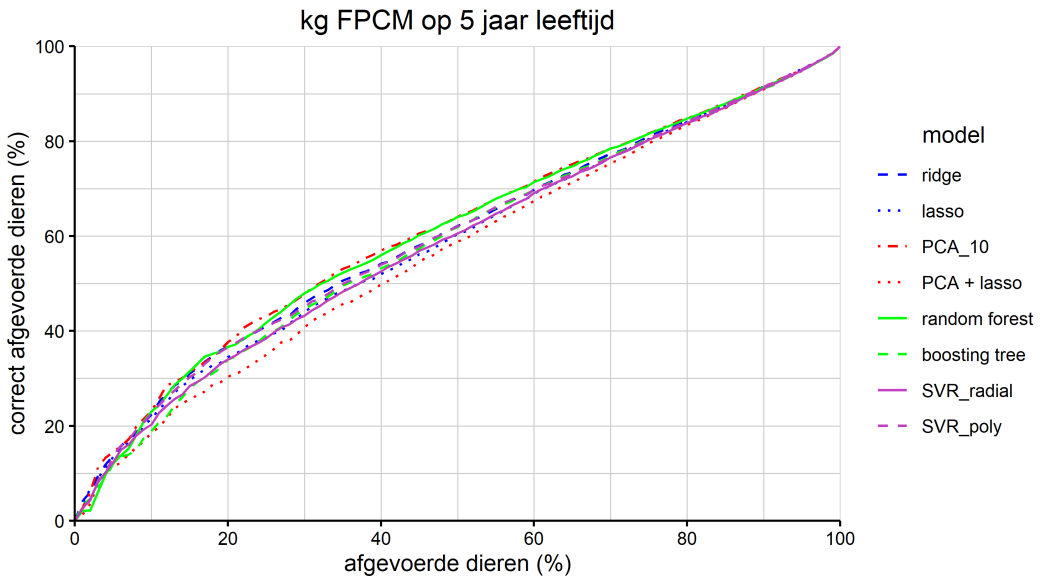


Figuur C.31: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde tussenkalftijden tussen de eerste en de tweede pariteit (EVI-modellen)

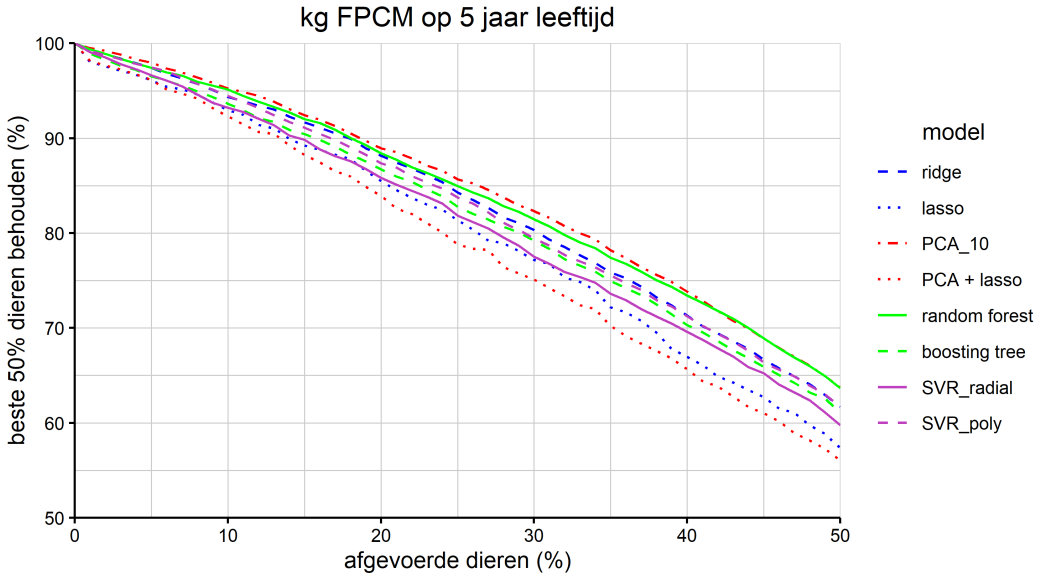


Figuur C.32: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde tussenkalftijden tussen de eerste en de tweede pariteit (EVI-modellen)

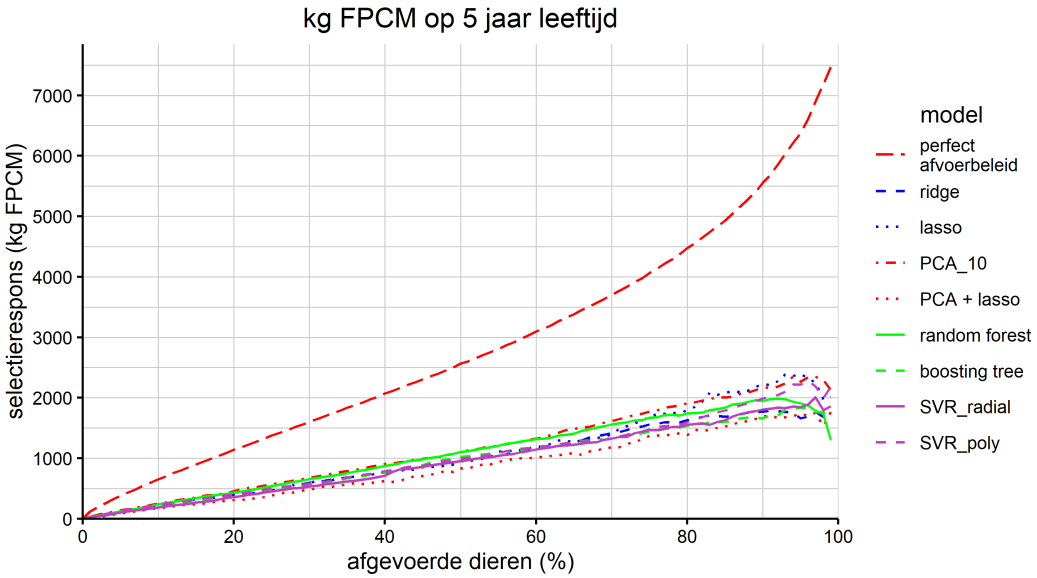
- **kg FPCM op 5 jaar leeftijd**



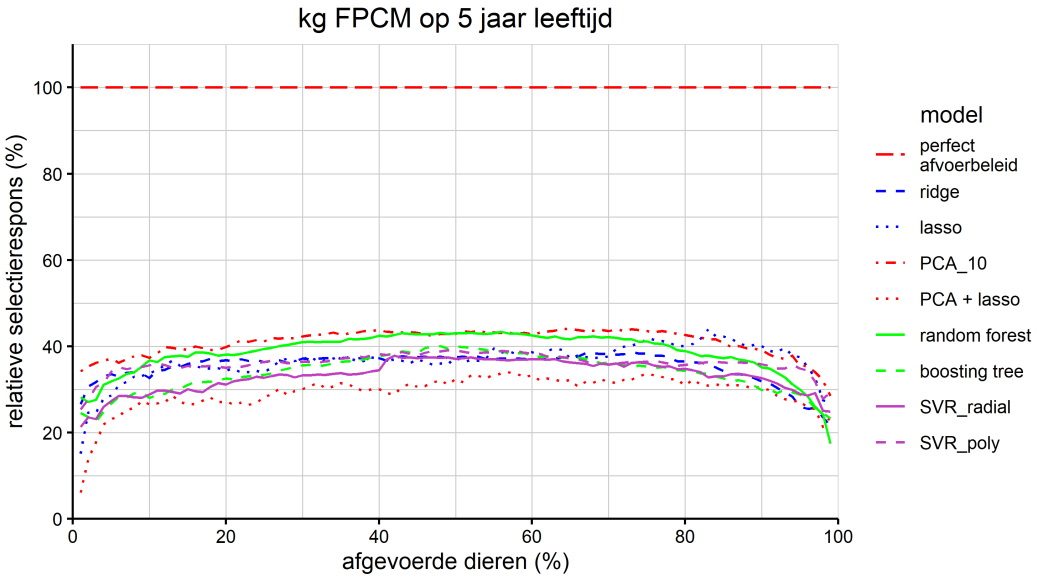
Figuur C.33: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde kg FPCM-producties op een leeftijd van 5 jaar (EVI-modellen)



Figuur C.34: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde kg FPCM-producties op een leeftijd van 5 jaar (EVI-modellen)

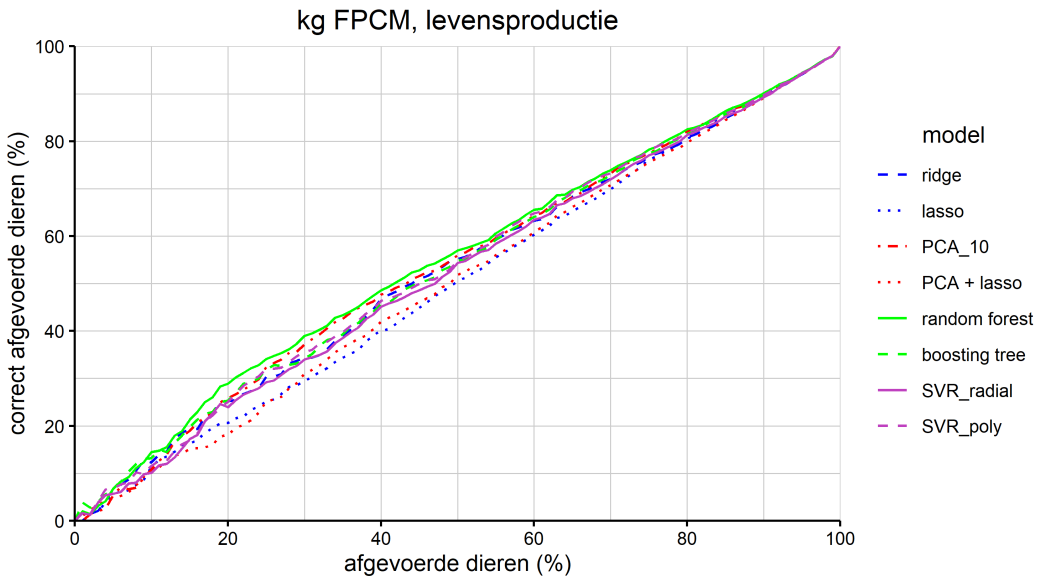


Figuur C.35: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde kg FPCM-producties op een leeftijd van 5 jaar (EVI-modellen)

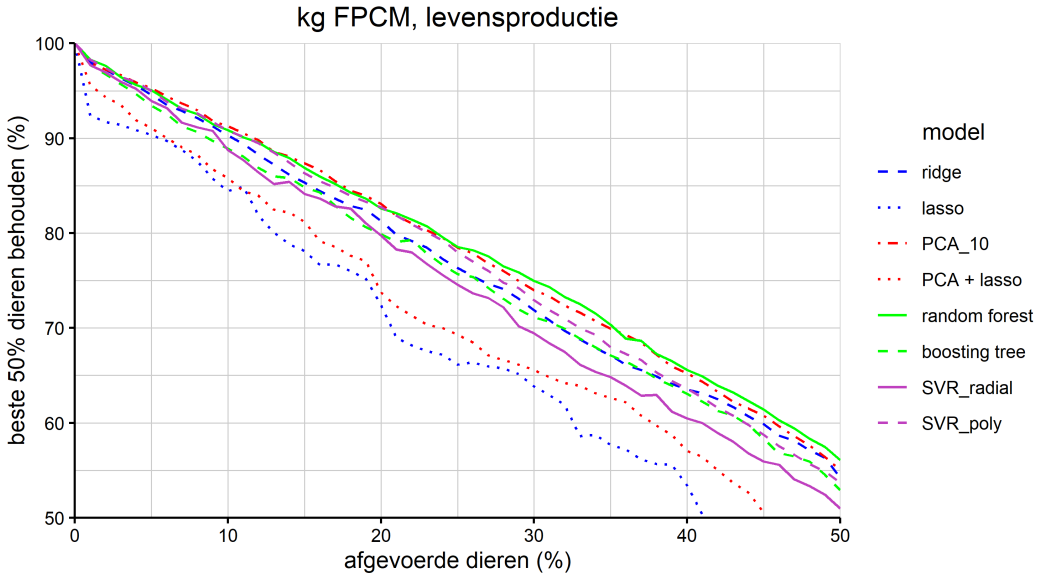


Figuur C.36: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde kg FPCM-producties op een leeftijd van 5 jaar (EVI-modellen)

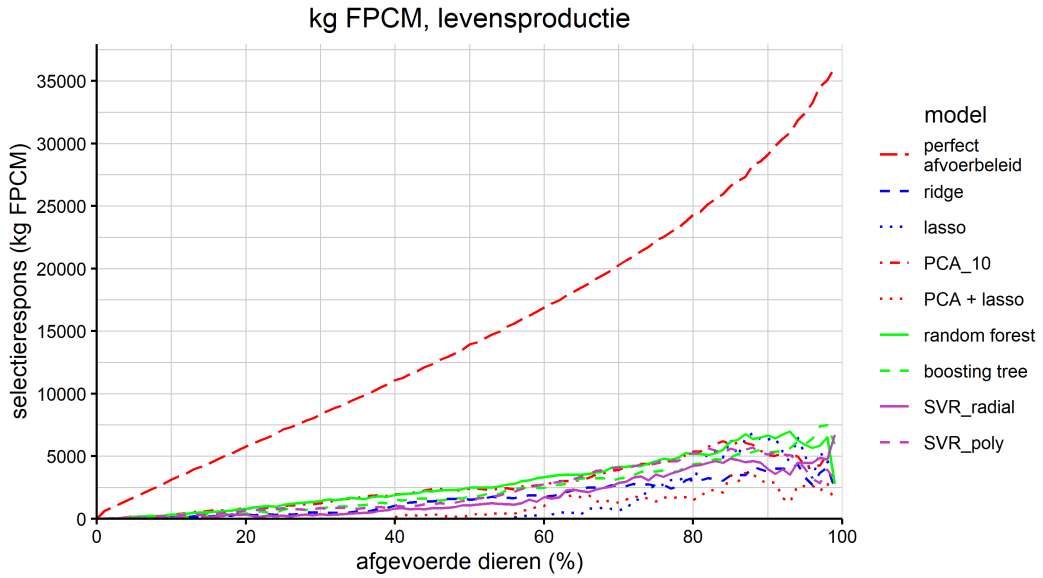
- **kg FPCM,levensproductie**



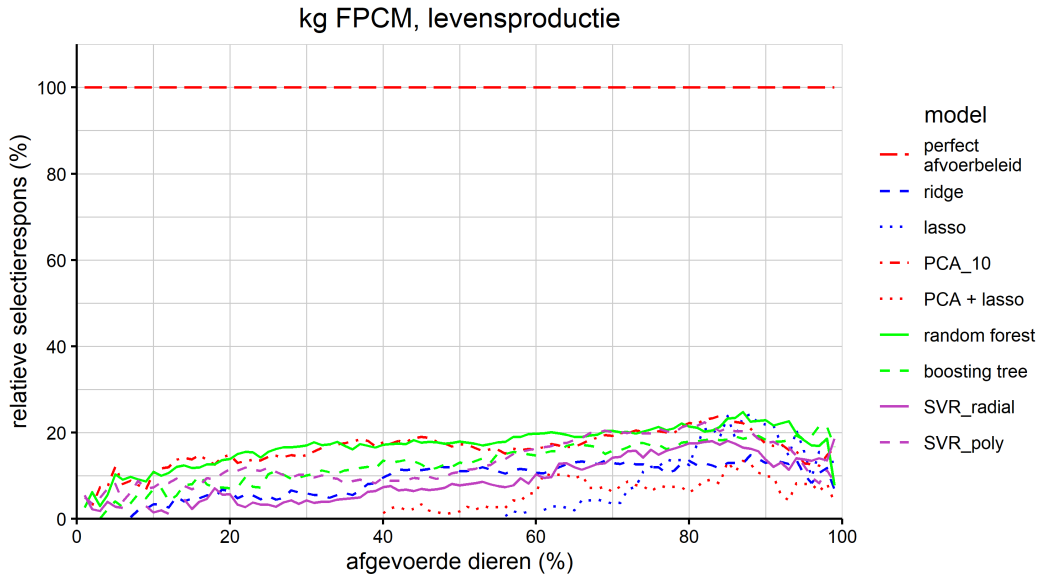
Figuur C.37: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM (EVI-modellen)



Figuur C.38: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM (EVI-modellen)



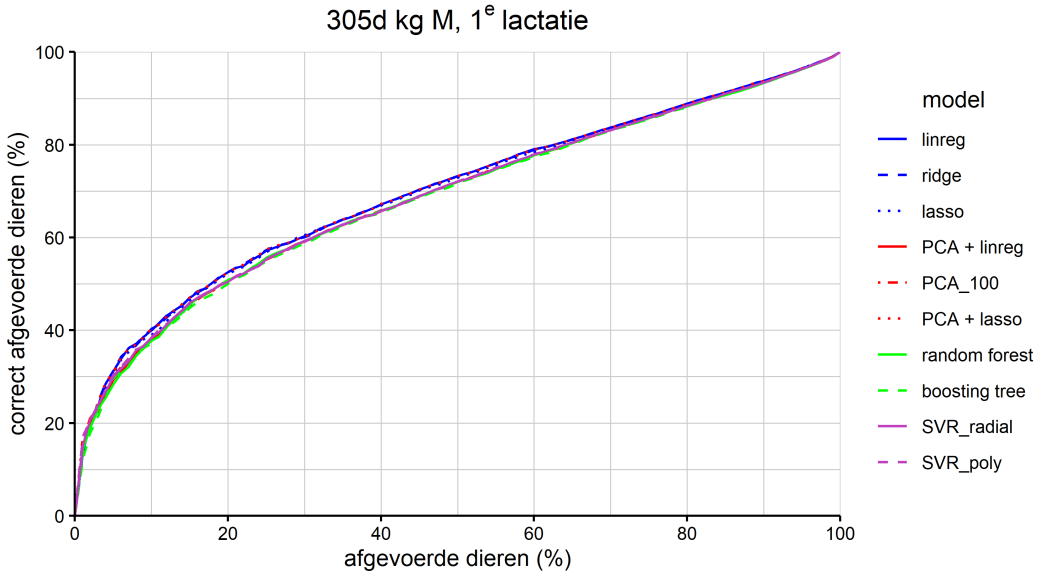
Figuur C.39: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM (EVI-modellen)



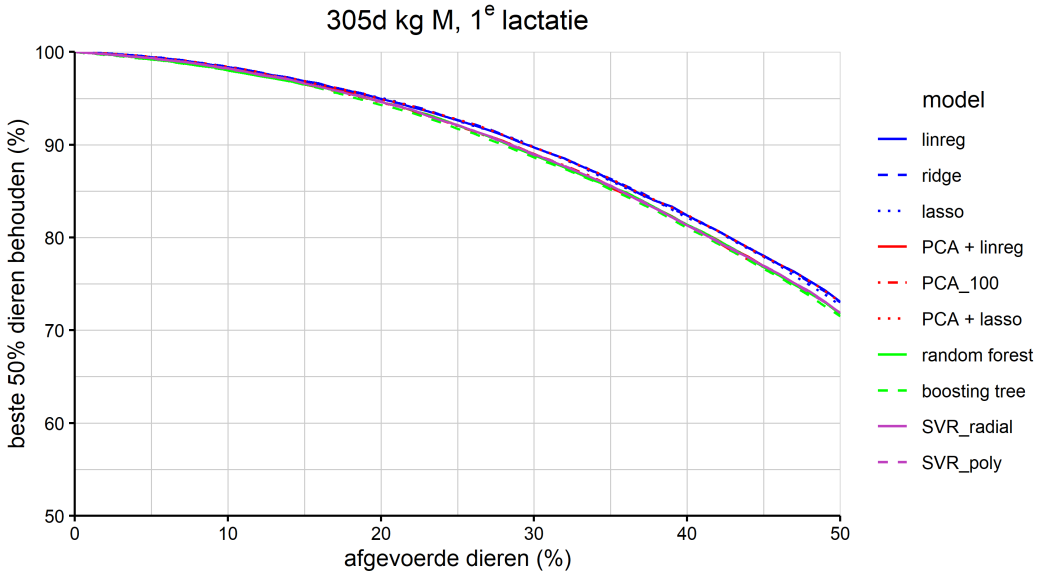
Figuur C.40: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM (EVI-modellen)

C.3 STIN: stacking

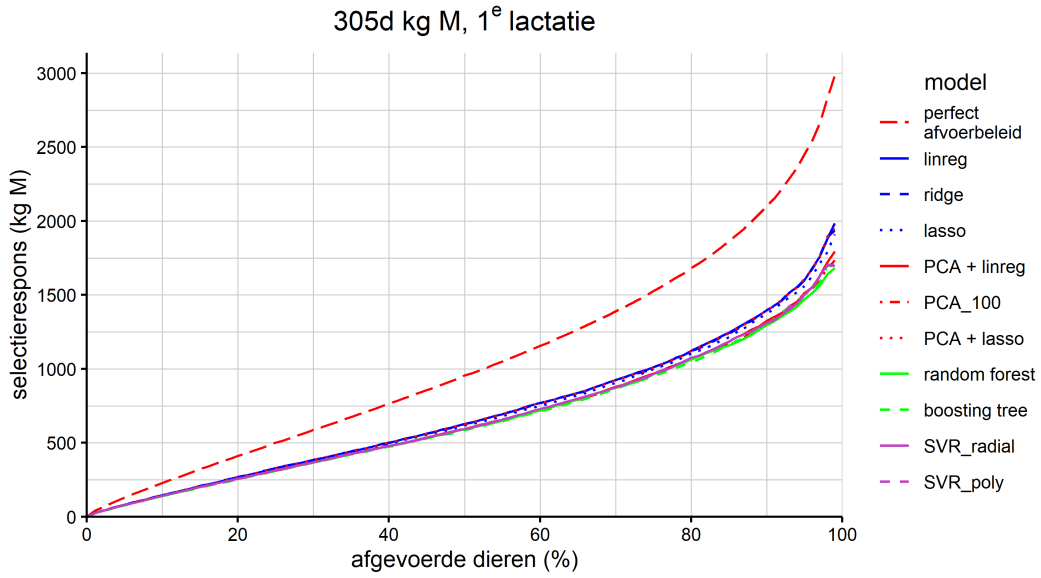
- **305d kg M, 1e lactatie**



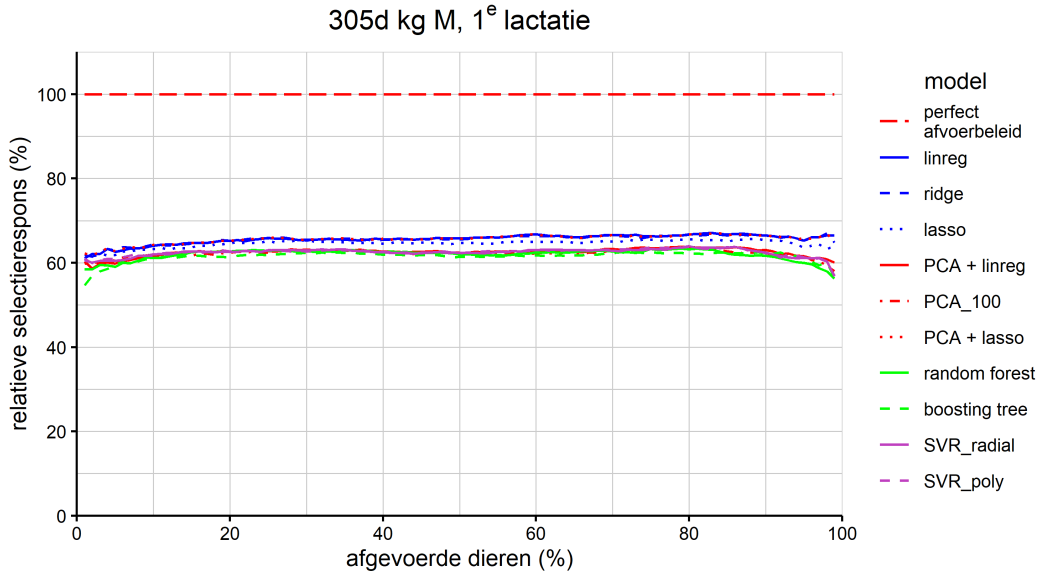
Figuur C.41: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg melk (STIN-modellen)



Figuur C.42: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg melk (STIN-modellen)

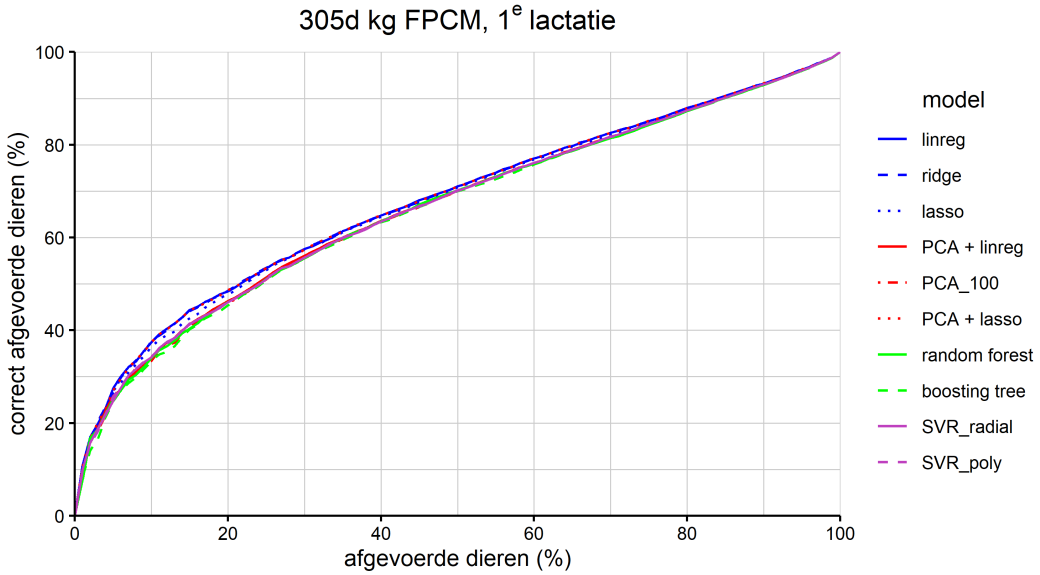


Figuur C.43: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg melk (STIN-modellen)

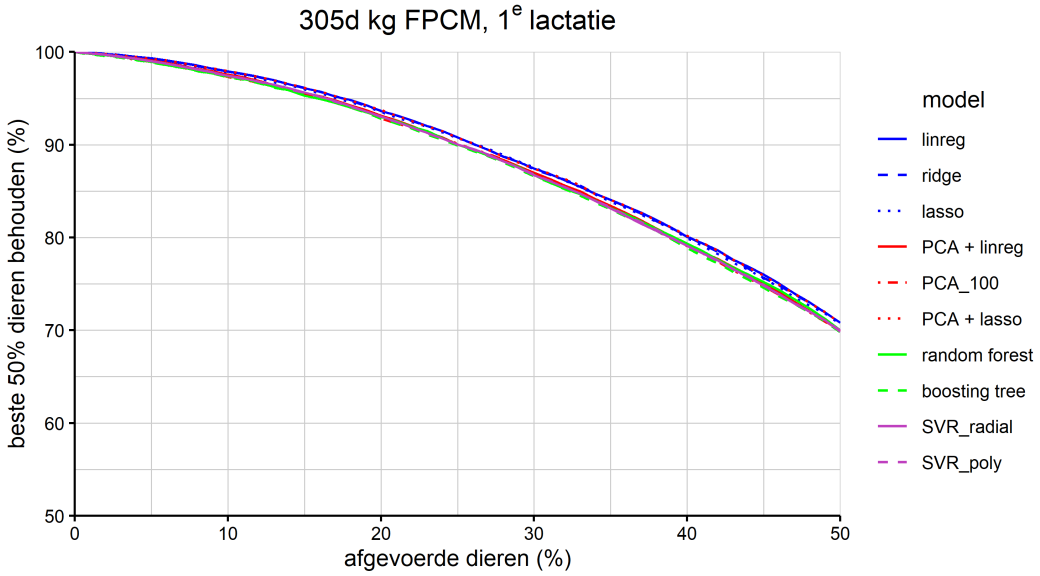


Figuur C.44: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg melk (STIN-modellen)

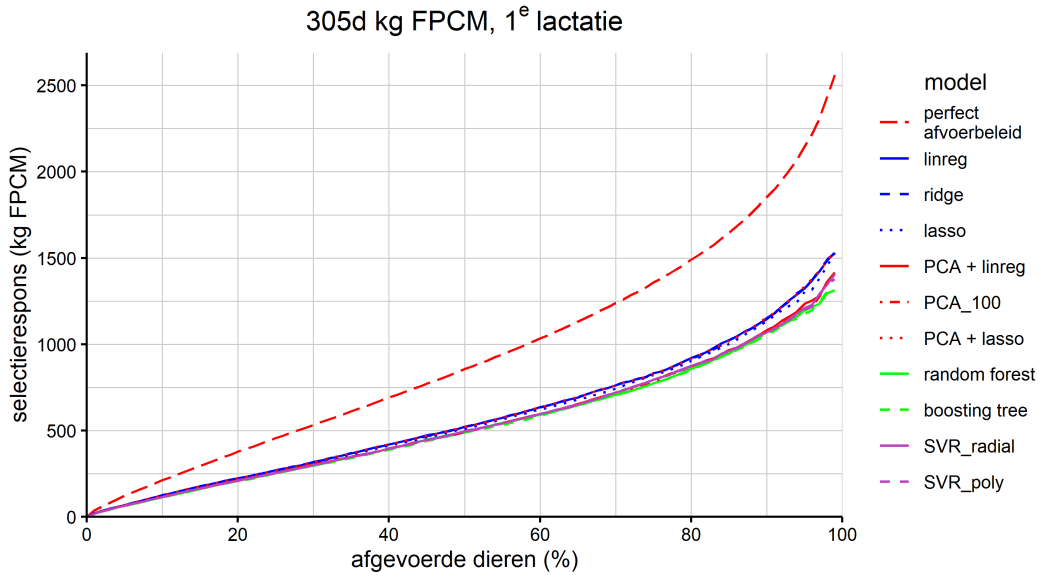
- **305d kg FPCM, 1e lactatie**



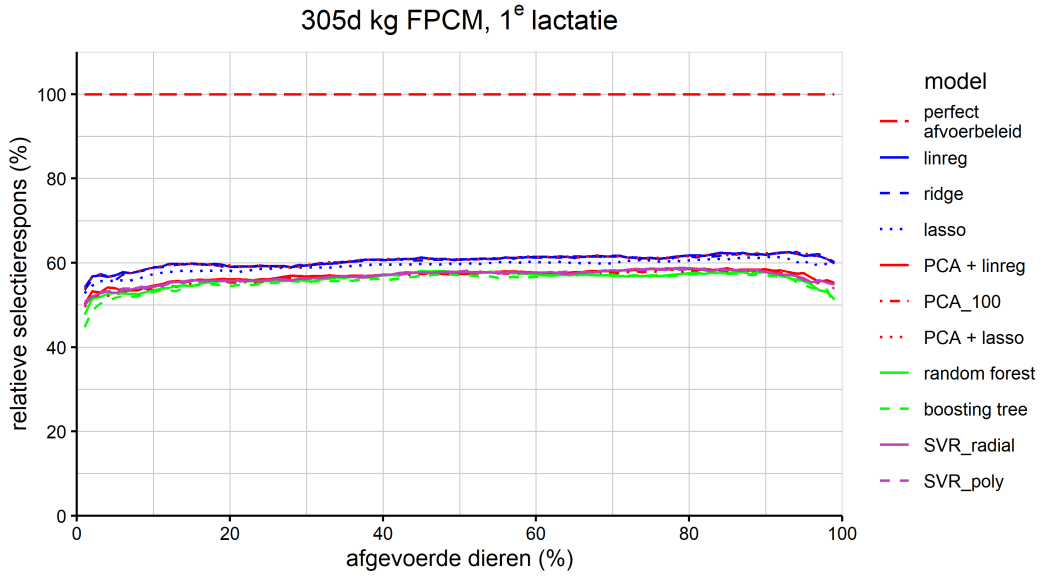
Figuur C.45: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg FPCM (STIN-modellen)



Figuur C.46: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg FPCM (STIN-modellen)

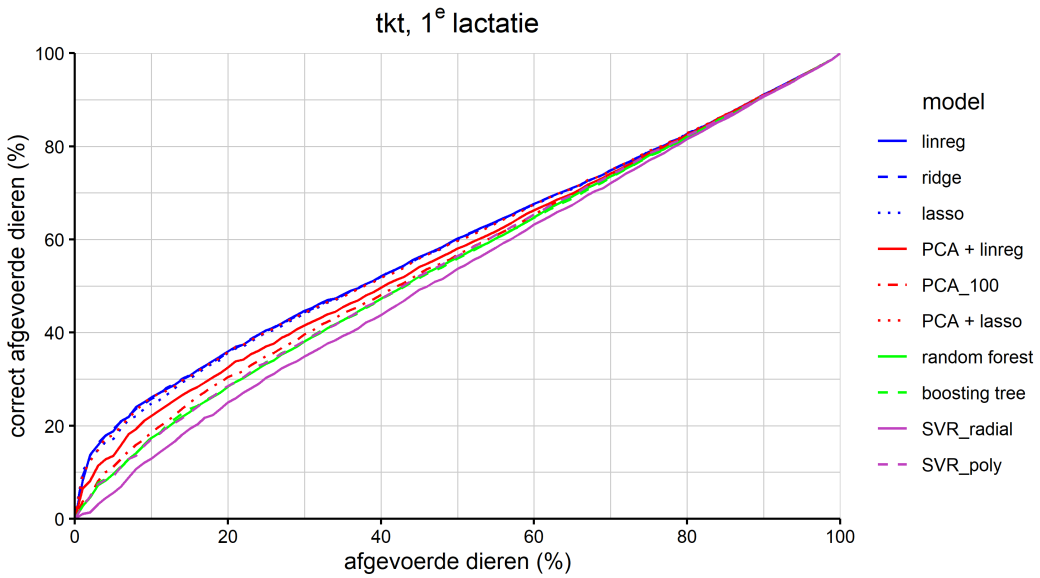


Figuur C.47: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg FPCM (STIN-modellen)

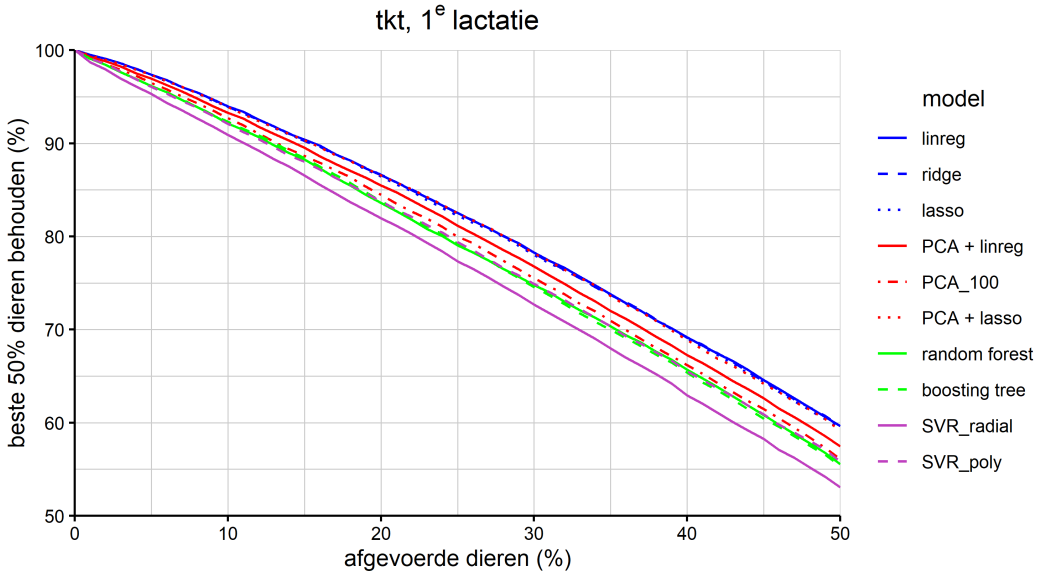


Figuur C.48: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde 305d-producties voor de eerste lactatie, uitgedrukt in kg FPCM (STIN-modellen)

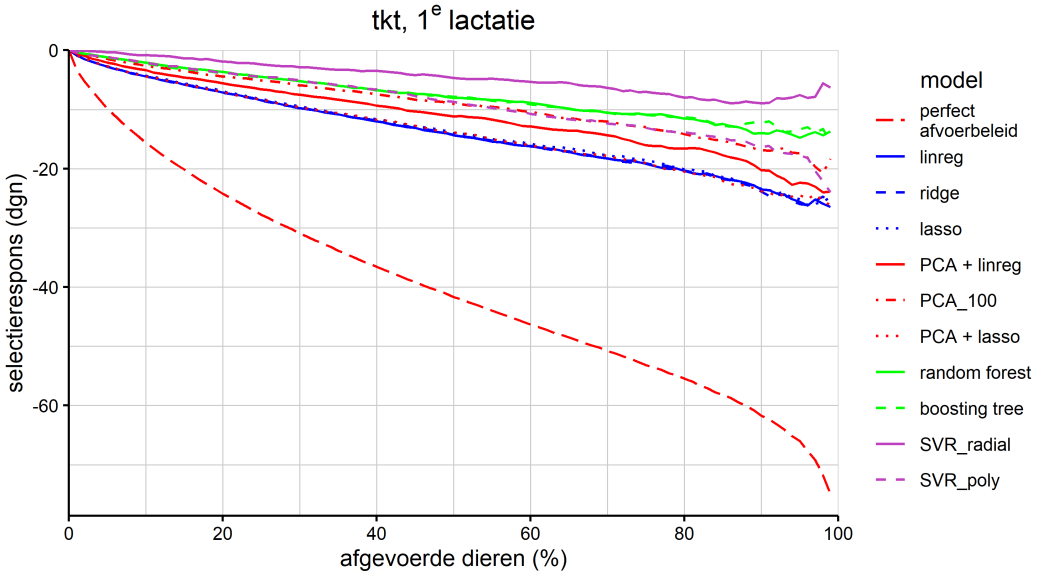
- **tkt, 1e lactatie**



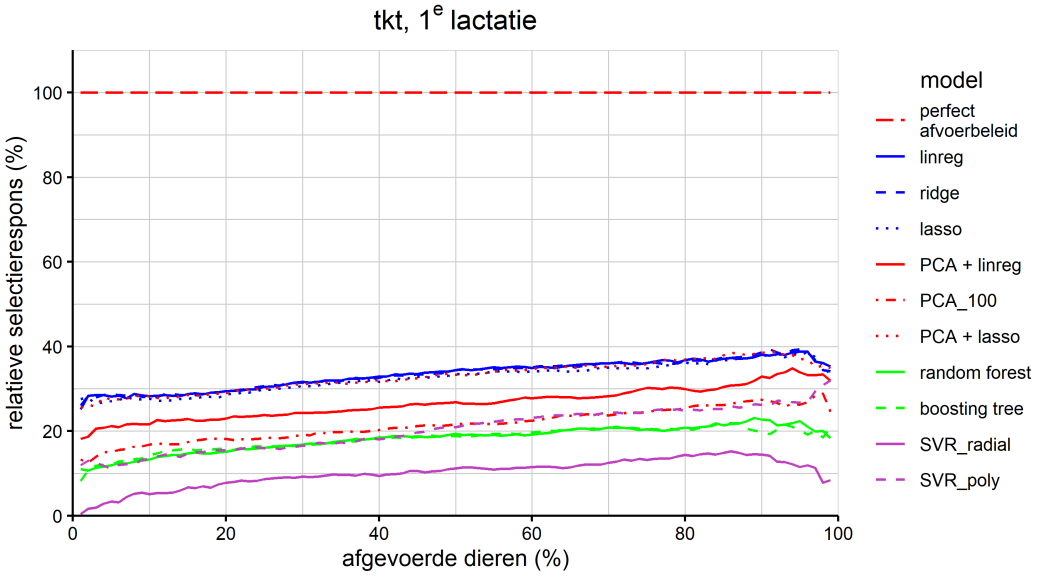
Figuur C.49: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde tussenkalftijden tussen de eerste en de tweede pariteit (STIN-modellen)



Figuur C.50: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde tussenkalftijden tussen de eerste en de tweede pariteit (STIN-modellen)

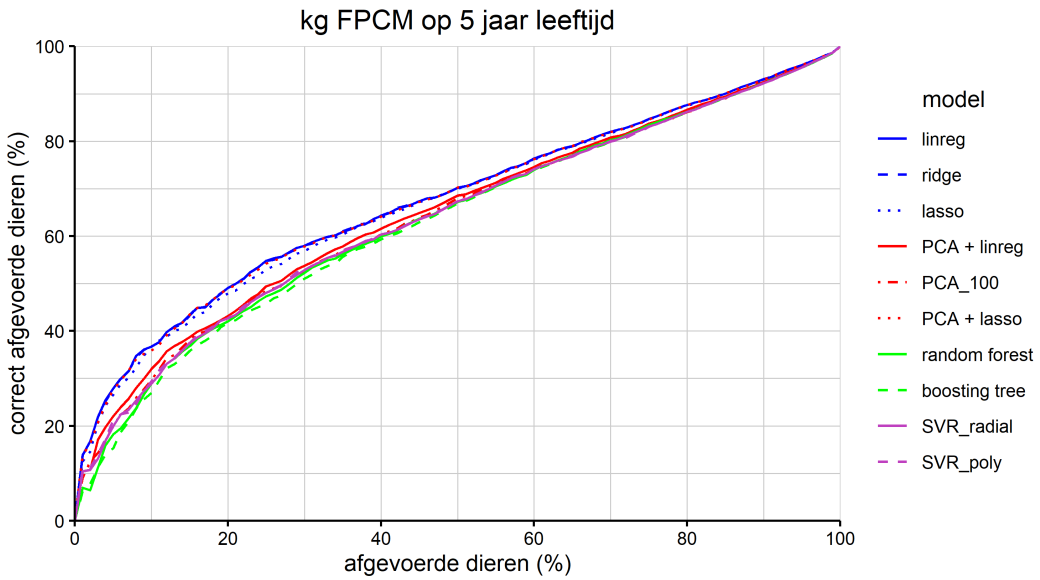


Figuur C.51: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde tussenkalftijden tussen de eerste en de tweede pariteit (STIN-modellen)

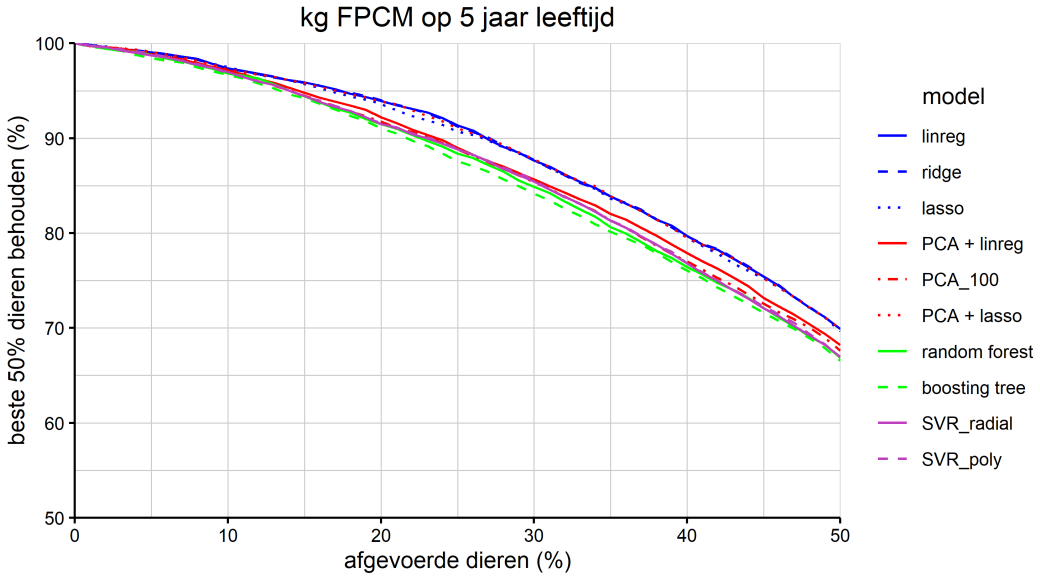


Figuur C.52: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde tussenkalftijden tussen de eerste en de tweede pariteit (STIN-modellen)

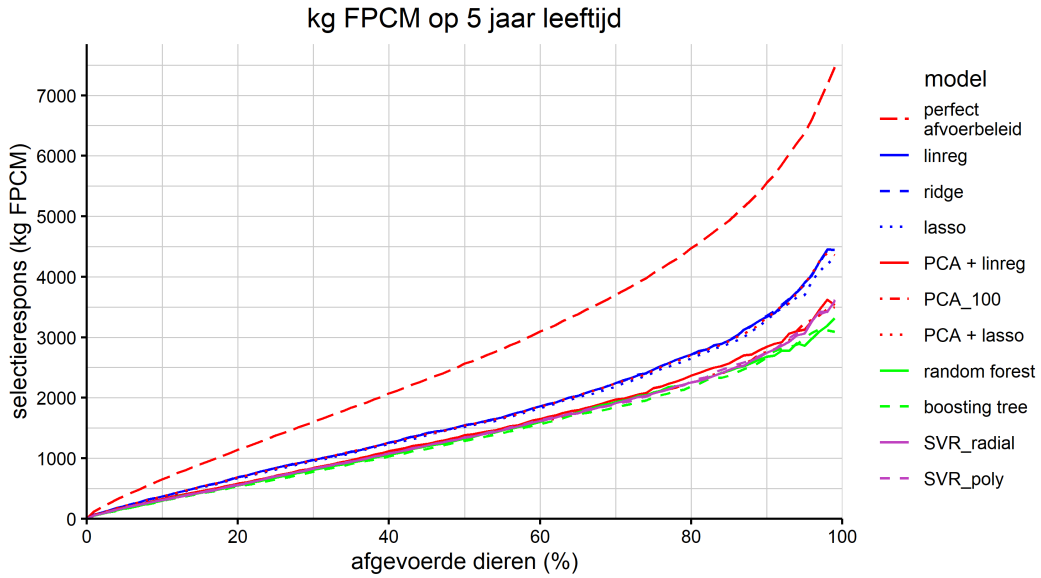
- **kg FPCM op 5 jaar leeftijd**



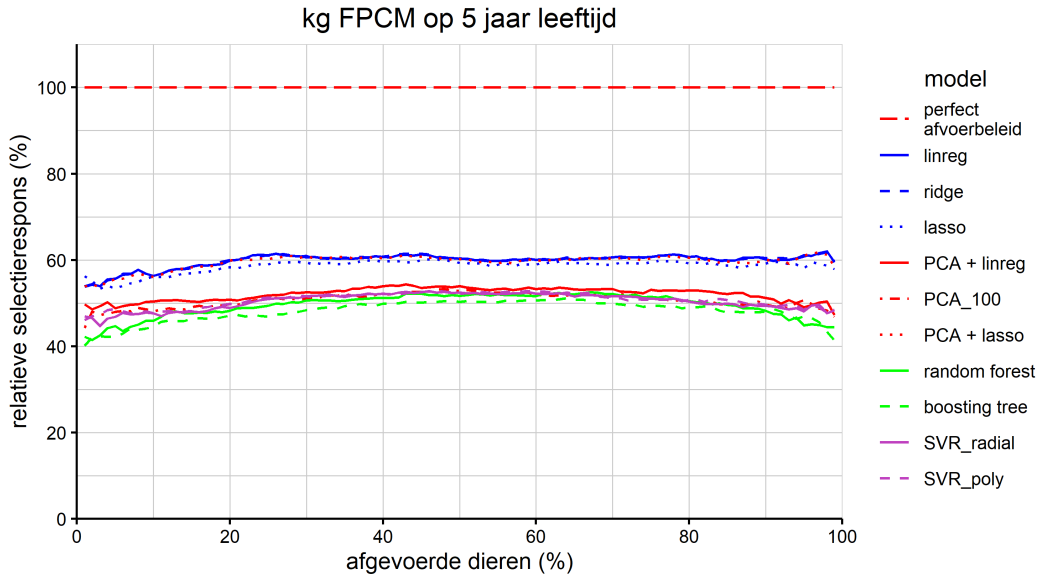
Figuur C.53: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde kg FPCM-producties op een leeftijd van 5 jaar (STIN-modellen)



Figuur C.54: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde kg FPCM-producties op een leeftijd van 5 jaar (STIN-modellen)

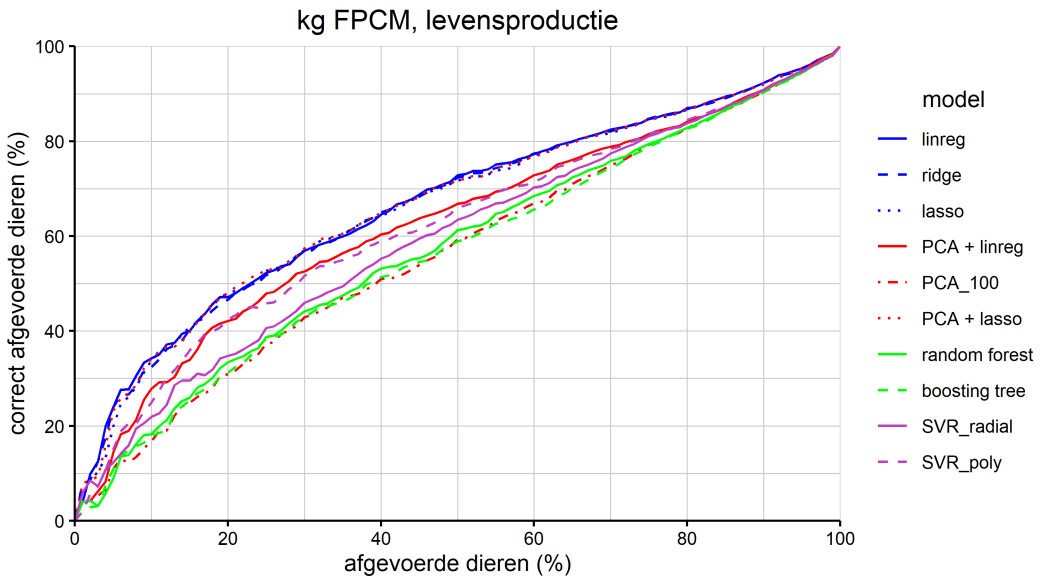


Figuur C.55: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde kg FPCM-producties op een leeftijd van 5 jaar (STIN-modellen)

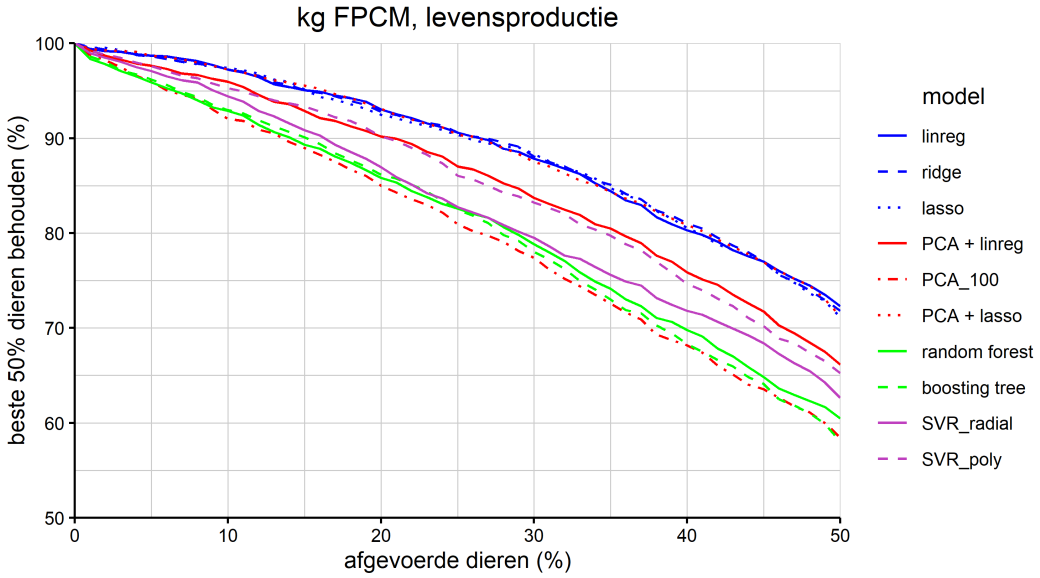


Figuur C.56: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde kg FPCM-producties op een leeftijd van 5 jaar (STIN-modellen)

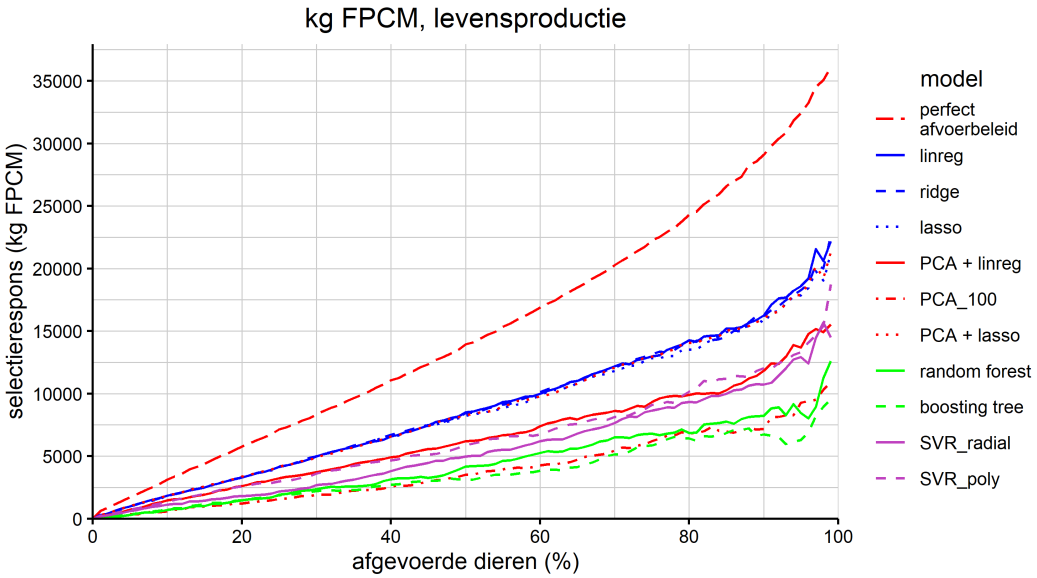
- **kg FPCM,levensproductie**



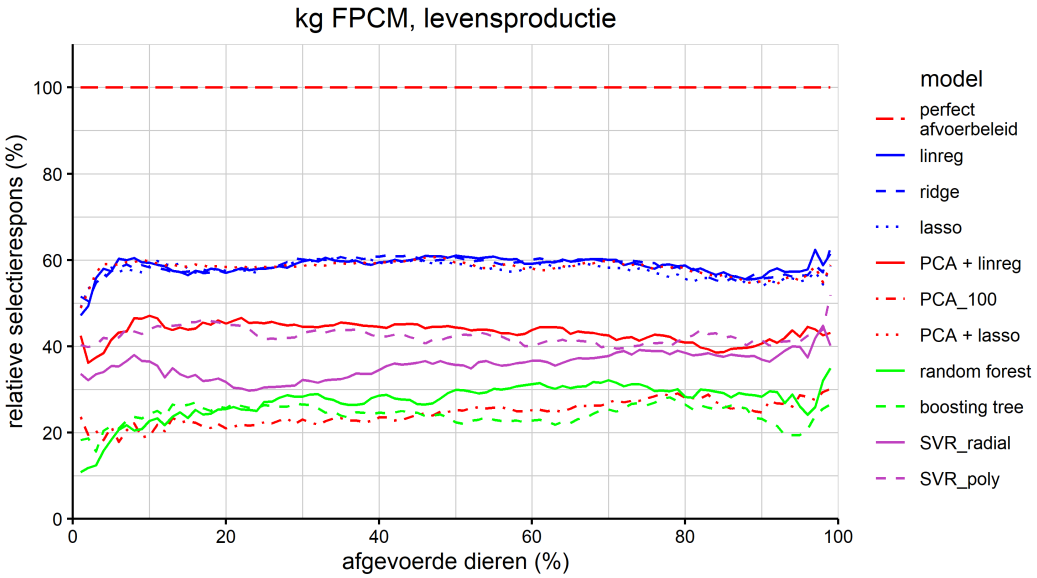
Figuur C.57: Percentage correct afgevoerde dieren bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM (STIN-modellen)



Figuur C.58: Percentage van de beste helft van de dieren dat wordt behouden bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM (STIN-modellen)



Figuur C.59: Selectierespons bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM (STIN-modellen)



Figuur C.60: Selectierespons, relatief uitgedrukt tegenover de maximaal haalbare selectierespons, bij een afvoerbeleid gebaseerd op voorspelde levensproducties, uitgedrukt in kg FPCM (STIN-modellen)